Pécsi Tudományegyetem
Bölcsészettudományi Kar
Nyelvtudományi Doktori Iskola
Alkalmazott Nyelvészeti Program

Aleksa Melita

A horvát sztenderd nyelv számítógépes morfológiai elemzésének alkalmazási lehetőségei német nyelvterületen

DOKTORI ÉRTEKEZÉS

Témavezető: Dr. Barics Ernő
egyetemi docens

Pécs
2008

Sveučilište u Pečuhu
Filozofski fakultet
Doktorska škola za lingvistiku
Smjer: primijenjena lingvistika

Melita Aleksa

Oblici primjene računalne morfološke analize hrvatskoga standardnog jezika na njemačkome govornom području

DOKTORSKA DISERTACIJA

Mentor:
doc.dr.sc. Ernest Barić

Pečuh
2008.

Iz razloga što sam tijekom stvaranja ovoga rada dobila potporu mnogih kolega, prijatelja i meni dragih osoba, na ovaj bih im način još jednom zahvalila na svemu pruženome.

Zahvaljujem prvenstveno svom mentoru, doc.dr.sc. Ernestu Bariću, na ukazanome povjerenju, potpori i konstruktivnim prijedlozima koji su uvelike pripomogli da ovaj rad dobije ovakav završni oblik.

Djelatnicima tvrtke MorphoLogic, posebno Attili Nováku i Lászlóu Tihanyiju, zahvaljujem na pomoći i mogućnosti korištenja i testiranja njihovih alata te na ukazanome povjerenju pri izradi HUMOR-a.

Posebnu zahvalu svakako zaslužuju prijatelji i kolege Snježana Babić, Mirela Berlančić, Zoltán Gotthardt, Nina Mance te Robert Wołosz na lekturi i korekturi napisanoga te što su svojim znanjem pridonijeli izradi ovoga rada.

Svakako veliku zahvalnost na pruženoj potpori i podršci zaslužuju moji roditelji, sestra, Domagoj Varga te Odsjek za germanistiku Filozofskoga fakulteta u Osijeku.

Naposljetku bih svoju zahvalu željela iskazati i Ministarstvu znanosti, obrazovanja i športa RH i Magyar Ösztöndíj Bizottságu na dodijeljenoj istraživačkoj stipendiji u ak. godini 2005./2006., bez koje ovaj rad ne bi mogao biti dovršen u ovome roku.

1. Uvod ... 9
1.1. Razlozi započinjanja izrade računalnoga morfološkog analizatora hrvatskoga standardnog jezika ... 9
1.2. Zašto računalni morfološki rječnik za govornike njemačkoga jezika? ... 11
1.3. Razlozi odabiranja procesa računalne morfološke analize ... 12

2. Analiza postojećih jezičnih materijala za učenje hrvatskoga jezika ... 13
2.1. Načelo analize nastavnih materijala ... 13
2.1.1. Korpusna analiza ... 16
2.1.2. Dtsearch ... 17
2.1.3. NSP ... 17
2.2. Analiza nastavnih materijala za učenje hrvatskoga jezika na njemačkome govornom području ... 21
2.3. Analiza jezičnih priručnika koji se rabe u nastavi hrvatskoga jezika ... 28
2.3.1. Glagoli ... 30
2.3.2. Pridjevi ... 37
2.3.3. Imenice ... 43
2.3.4. Zamjenice i brojevi ... 48
2.3.5. Analiza Rječnika hrvatskoga jezika ... 49

3. Morfološka analiza u službi izrade morfološkoga rječnika ... 52
3.1. Načelo izrade morfološkoga rječnika hrvatskoga standardnog jezika ... 52
3.2. Izbor jezične inačice za izradu morfološkoga rječnika ... 55

4. HUMOR kao jedno od sredstava za računalnu morfološku analizu ... 59
4.1. Način rada HUMOR-a ... 60
4.2. Struktura morfološkoga analizatora ... 65
4.3. Definiranje pojmova ... 70

5. Prikaz postojećih verzija morfoloških analizatora za aglutinativne, fleksijske i visokofleksijske jezike ... 75

6. Izrada hrvatske inačice morfološkoga analizatora ... 80
6.1. Problematika jezične prošlosti i jezične politike pri postupku morfološke analize ... 80
6.2. Problematika jezične inačice pri izradi morfološkoga analizatora ... 88
6.3. Testni korpus ... 90

7. Primjena morfološkoga analizatora za učenje i poučavanje hrvatskoga standardnog jezika ... 97
7.1. Problematika obrade promjenjivih vrsta riječi i primijenjena rješenja ... 99
7.2. Imenice ... 99
7.2.1. Imenska paradigma ... 100
7.2.2. Problematika ... 106
7.3. Pridjevi ... 108
7.3.1. Stupnjevanje pridjeva ... 114
7.3.2. Jezične nedoumice ... 115
7.3.3. Pridjevska paradigma u računalnoj obradi hrvatskoga jezika u usporedbi s pridjevskim paradigmama u drugim jezicima ... 116
7.4. Glagoli ... 118

7.4.1. Glagolska paradigma 118
7.4.2. Glagolska paradigma unutar morfološkoga analizatora 121
7.4.3. Problematika 125

8. Računalni morfološki rječnik hrvatskoga standardnog jezika 127
8.1. Struktura rječnika 127
8.2. Tehnička rješenja 128
8.3. Mogućnosti morfološkoga rječnika 132
8.4. Jezični problemi i rješenja koja su primijenjena pri izradi računalnoga morfološkog rječnika 135
8.4.1. Problematika padeža 135
8.4.2. Problematika rodova 141

9. Ostali oblici primjene morfološke analize 149
9.1. Jezični alati za učenje i poučavanje jezika 151
9.1.1. Jednojezični i dvojezični računalni rječnici 153
9.1.2. Jezični alati za učenje mađarskoga jezika 157
9.2. Programi za strojno prevođenje 158

10. Završna riječ i perspektive 161

11. Összefoglaló 165

12. Zusammenfassung 189

Literatura 211

Prilozi 217

1. Uvod

Već je iz samoga naslova rada vidljivo da se on tematski usredotočuje na jednu jezičnu inačicu, ali istodobno obuhvaća dvije jezične skupine s procesom morfološke analize kao svojim polazištem. Međutim, prije početka obrade same teme smatram nužnim podrobnije objasniti razloge njezina izbora.

1.1. Razlozi započinjanja izrade računalnoga morfološkog analizatora hrvatskoga standardnog jezika

Nakon objavljivanja *Zajedničkog europskog referentnog okvira za jezike*[1] Vijeća Europe pojavila se na tržištu povećana potreba za izradom nastavnih materijala za učenje i poučavanje hrvatskoga kao drugoga ili stranoga jezika, a koji bi pratio navedeni okvir. Interpretacijom spomenutih smjernica posebice se naglašavao pojam tradicionalnih gramatičkih razina kao sastavnica jezične sposobnosti (Jelaska 2005a: 22). Tako primjerice pri razmatranju pojma komunikacijske jezične kompetencije u Europskome jezičnom okviru Jelaska zaključuje da bi se „trebalo odlučiti na navođenje gramatičkih sposobnosti, koje se sastoje od gramatičkih vještina (tj. primjene), kao suodnosne razine vještinama razgovornoga sporazumijevanja. (…) Jezična bi sposobnost uključivala gramatičku (s fonetskom, fonološkom, morfološkom, sintaktičkom semantičkom i pragmatičkom), komunikacijsku i društvenu sposobnost“ (Jelaska 2005a: 23). Ako se promotri nekoliko projekata pokrenutih u tom pravcu u Hrvatskoj i u inozemstvu, primjerice projekt *Hrvatski kao strani jezik: razvojna gramatika i rječnik* (MZOŠ 130738)[2], koji je trajao od 14. ožujka 2001. do 22. kolovoza 2002. godine, i njegov nastavak, projekt *Hrvatski kao drugi i strani jezik* (MZOŠ 0130438)[3], može se zaključiti da se već od 2001. godine intenzivno bavilo problemom gramatičke kompetencije[4]. Ishodom prvoga projekta navodi se nadopuna gramatike hrvatskoga jezika na hrvatskome i engleskome jeziku te izdavanje skripte za

[1] Council of Europe (2001). *Common European Framework of Reference for Languages*: *leraning, teaching, assesment*. Cambridge: Cambrige University Press, dalje u tekstu ZEROJ
[2] http://www.croatiana.org/croatiana--projekti-hrvatskikaostrani.htm, 4. veljače 2008.
[3] http://www.croatiana.org/croatiana--projekti-hrvatskikaodrugi.htm, 4. veljače 2008.
[4] Osim navedenih dvaju projekata, prema mojim saznanjima, pokrenuto je i nekoliko ostalih projekata u Hrvatskoj i u inozemstvu: projekt *Hrvatski za strance* voditelja I. Pranjkovića, projekt *Tempus* Europske unije, *Projekt višejezičnoga razvoja* Instuta Ludwig Boltzmann iz Beča te neki slovenski projekti (Cvikić-Jelaska 2005a:129). S obzirom na to da su za ovaj rad relevantniji rezultati projekata *Hrvatski kao strani jezik: razvojna gramatika i rječnik* i *Hrvatski kao drugi i strani jezik*, u radu će se podrobnije razmotriti i navesti njihovi ciljevi te dosadašnji rezultati.

potrebe Sveučilišne škole hrvatskoga jezika i kulture. Prikupljena je i građa za izradu temeljnoga morfološkog rječnika za učenje hrvatskoga jezika te je izrađen popisnik sa 7000 najčešćih riječi hrvatskoga jezika dobivenih iz različitih izvora[5]. Drugim je projektom pak nastao *Morfološki rječnik MORKO*[6] te *Glagolnica* koja je prethodila izdanju *Hrvatski glagoli – oblici*[7].

Kao što se može zaključiti, jednim od ciljeva projekata postavila se i izrada materijala koji pomažu razvijanju gramatičke sposobnosti kod učenika hrvatskoga kao drugoga ili stranoga jezika[8]. Navedeni materijali naime sadržavaju odabranu gramatičku građu, određeni broj riječi, odnosno pri njihovoj se izradi koristilo strogo odabranim i kontroliranim vokabularom. Zaključak je istraživanja o djelotvornosti gramatičkoga poučavanja (Novak-Milić 2005: 353) da se „strance treba izravno poučavati gramatici, ali na primjeren način uzevši u obzir i brojne druge okolnosti i načine". Osim toga predvači hrvatskoga kao stranoga jezika mišljenja su da „postojeće gramatike za izvorne govornike ne odgovaraju ni potrebama predavača ni potrebama učenika" (Cvikić 2005a: 320) te ističu nužnost izdavanja „većeg broja priručnika pisanih za hrvatski kao strani jezik: gramatika i gramatičkih vježbenica, udžbenika za opći i stručni jezik, prilagođenih lektira, rječnika i slično" (Cvikić 2005a: 320). Uzimajući navedene postavke u obzir, logičnim se nastavkom navedenih dostignuća u ovome radu, sukladno navedenim ciljevima, postavilo razvijanje materijala u svrhu svladavanja hrvatskoga kao drugoga ili stranoga jezika i na višim razinama. Kao polazište za njihovu izradu postavio se računalni morfološki analizator hrvatskoga standardnog jezika s leksičkom bazom od približno 60 000 hrvatskih leksema.

Dva su krajnja cilja primjene računalnoga morfološkog analizatora: Prvi je omogućiti učenicima hrvatskoga kao drugoga ili stranoga jezika prikaz konkretnih deklinacijskih i

[5] http://www.croatiana.org/croatiana--projekti-hrvatskikaodrugi.htm, 4. veljače 2008.

[6] Jelaska, Z., L. Cvikić (2003) *Morko - Hrvatski višejezični morfološki rječnik* Zagreb: Sveučilišna škola hrvatskoga jezika i kulture; ovo izdanje dostupno je samo polaznicima Sveučilišne škola hrvatskoga jezika, nije dostupno za širu uporabu pa izdanje nisam mogla podrobnije pogledati i analizirati.

[7] Jelaska, Z. (2003) *Hrvatski glagoli - oblici*, Zagreb: Sveučilišna škola hrvatskoga jezika i kulture; ovo izdanje dostupno je samo polaznicima Sveučilišne škola hrvatskoga jezika, nije dostupno za širu uporabu pa izdanje nisam mogla podrobnije pogledati i analizirati.

[8] Prema riječima autora „osmišljen je nov način poučavanja hrvatskoga glagolskoga sustava i izrađena nova podjela na glagolske vrste prikladnija učenju hrvatskoga kao stranoga jezika. Iz rječničke je građe odabrano 5000 glagola za koje su ispisane paradigme i ostali relevantni gramatički podatci. Istraživanjem gramatičkoga razvoja odabrana je građa za početno učenje hrvatskoga jezika (P1A i P1B), izrađena je gramatička vježbenica sa strogo kontroliranim rječnikom, a prvi dio gramatike hrvatskoga jezika na engleskome jeziku s obiljem primjera dopunjen je novim podatcima. Uporaba vježbenice i gramatike u obliku skripta provjerena je na nastavi u Sveučilišnoj školi hrvatskoga jezika i kulture. U trećoj godini rada nastavljen je rad na istraživanju razvojne gramatike hrvatskoga jezika, odabrana je gramatička građa za višu, početnu, nižu i srednju razinu učenja hrvatskoga kao nematerinskoga jezika (P2, P3, S1). Izrađena je gramatička vježbenica za ove razine učenja hrvatskoga, a onda je provjerena u nastavi u Sveučilišnoj školi hrvatskoga jezika i kulture"(http://www.croatiana.org/croatiana--projekti-hrvatskikaodrugi.htm, 4. veljače 2008.).

konjugacijskih paradigmi u obliku jednoga računalnoga morfološkog rječnika koje se temelje kako na stručnim navodima iz literature, tako i na korpusnoj analizi, odnosno uporabnome jeziku. Rječnik ne sadržava samo odabrane riječi, nego osim lema iz Anićeva *Velikog rječnika hrvatskoga jezika*[9], pokriva i one leme koje nisu navedene u spomenutome rječniku, a za koje se uz pomoć korpusne analize ispostavilo da se ubrajaju među češće pojavnice. Drugim se ciljem izrade morfološkoga analizatora postavila izrada ostalih pomagala za učenje i poučavanje hrvatskoga kao drugoga ili stranoga jezika u obliku računalnih rječnika i prevoditeljskih pomagala, odnosno programa za strojno prevođenje s hrvatskoga i na hrvatski jezik.

Iz razloga što je tema ove disertacije opis samoga postupka i naznaka problema koji su se pojavili pri izradi računalnoga morfološkog analizatora hrvatskoga standardnog jezika, te plan izrade morfološkoga rječnika za govornike njemačkoga jezika, podastrijet će se samo teoretski opis izrade ostalih spomenutih pomagala.

1.2. Zašto računalni morfološki rječnik za govornike njemačkoga jezika?

Razlozi odabiranja govornika njemačkoga leže u broju hrvatskih iseljenika i njihovih potomaka u svijetu (Prilog 1). Ako se podrobnije analiziraju podatci iz izvješća Ministarstva vanjskih poslova i europskih integracija Republike Hrvatske, vidljivo je da se najveći broj hrvatskih iseljenika nalazi na engleskome i njemačkome govornom području. Ako se uzme u obzir činjenica da je iseljavanje Hrvata na njemačko govorno područje započelo 1918. godine[10], i pretpostavka da govornici kojima materinski jezik nije hrvatski (primjerice većina učenika pripadnika mađarske nacionalne manjine u Republici Hrvatskoj) imaju poteškoća prilikom komuniciranja na hrvatskome jeziku[11], potreba izrade dodatnih materijala za učenje hrvatskoga kao drugoga ili stranoga jezika i na višim

[9] Anić, V. (2000). *Veliki rječnik hrvatskoga jezika*. Zagreb: Novi Liber

[10] http://www.mvpei.hr/hmiu/tekst.asp?q=osi001, 4. veljače 2008.

[11] Ova se izjava temelji na anketi provedenoj među profesorima hrvatskoga jezika u Prosvjetno-kulturnom centru Mađara u Republici Hrvatskoj u Osijeku i sociolingvističkom istraživanju koje sam 2004. godine provela među učenicima navedene škole. Rezultati su istraživanja pokazali da su učenici pripadnici mađarske nacionalne manjine u omjeru 85% dvojezični govornici s izrazitom dominacijom mađarskoga jezika, te da ih se svega 24% u svakodnevnoj interakciji koristi hrvatskim jezikom (Aleksa 2007a). Prema subjektivnoj procjeni profesora hrvatskoga jezika u PKCM-u, hrvatska bi se jezična kompetencija većine mađarskih učenika mogla odrediti kao razina B1 i B2 Europskoga referentnog okvira. Za određivanje njihove stvarne jezične kompetencije te definiranje poteškoća i problema u svladavanju hrvatskoga jezika i postizanju viših razina znanja, potrebna su dodatna istraživanja.

razinama može se smatrati opravdanom. Osim toga, iz razloga što materijali slični ovomu projektu za govornike engleskoga jezika već postoje[12], dok prema mojim saznanjima radovi slične tematike i opsega za govornike njemačkoga jezika još nisu objavljeni, pri nastanku ovoga rada pristupilo se izradi hrvatskoga računalnog morfološkog rječnika za učenike hrvatskoga kao drugoga ili stranoga jezika.

1.3. Razlozi odabiranja procesa računalne morfološke analize

Postoje dva razloga odabiranja procesa morfološke analize kao polazišta za izradu računalnoga morfološkog rječnika. Prvi razlog leži u načinu rada morfološkoga parsera HUMOR-a[13], odnosno morfološkoj analizi kao njegovom osnovicom, dok je drugi razlog činjenica da morfološki rječnik koji je temeljen na morfološkoj analizi, odnosno korpusnoj analizi sadrži paradigme koje se temelje na jezičnoj uporabi, što se ne može uvijek reći za navode koji se temelje na morfološkoj sintezi, odnosno navodima iz gramatika i jezičnih priručnika (poglavlje 2).

[12] Primjerice sljedeća izdanja:
Jelaska, Z. (2003) *Basic Croatian Grammar I: Sounds, Forms, Word-Types***.** Zagreb: Sveučilišna škola hrvatskoga jezika i kulture
Grubišić, V. (1995) *Croatian Grammar*. Zagreb: Hrvatska sveučilišna naklada

[13] HUMOR, kratica za High-Speed Unification based Morphology, jest morfološki parser koji je razvila tvrtka MorphoLogic za morfološku analizu tekstova te se danas rabi kao osnovica za razne programe kojima je temelj morfološka analiza. Više o samome programu nalazi se u poglavlju 4 ovoga rada. U poglavljima 6 i 9 slijedi detaljan opis programa u svrhu implementacije u sustave koji podupiru učenje i poučavanje stranih jezika.

2. Analiza postojećih jezičnih materijala za učenje hrvatskoga jezika

Polazeći od već citirane izjave predavača hrvatskoga kao stranoga jezika da postojeće gramatike za izvorne govornike ne odgovaraju potrebama učenika te da je potreban veći broj priručnika pisanih za hrvatski kao strani jezik: gramatika i gramatičkih vježbenica, udžbenika za opći i stručni jezik te rječnika (Cvikić 2005a: 320), u ovome će se poglavlju dati pregled i analiza nastavnih materijala koji se najčešće rabe u nastavi hrvatskoga kao drugoga i stranoga jezika na razinama B1 i B2. Analiza je materijala osim toga proširena i na gramatičke priručnike koji se većinom rabe u nastavi kako hrvatskoga kao materinskoga, tako i kao drugoga ili stranoga jezika. Osim publiciranih materijala u tiskanome obliku podastrijet će se prikaz jezičnih materijala dostupnih i na elektronskim medijima.

S obzirom na to da je cilj ovoga rada poduzimanje uvodnih radnji za izradu hrvatskoga računalnoga morfološkog rječnika za govornike njemačkoga jezika, namjera je analize nastavnih materijala prikaz leksika kojim se koristi u materijalima za poučavanje hrvatskoga kao drugoga i stranoga jezika govornicima njemačkoga te točnosti, jasnosti i uporabnosti gramatičkih objašnjenja. U sljedećim odlomcima slijedi detaljniji prikaz samoga procesa te postavljenih ciljeva analize.

2.1. Načelo analize nastavnih materijala

Osnovnim polazištem za analizu leksika nastavnih materijala, unutar ciljeva ovoga rada, postavljen je opis razina komunikacijske jezične kompetencije iz ZEROJ-a, odnosno unutar njega definirane leksičke i gramatičke sposobnosti unutar jezične kompetencije. Prema riječima Jelaske (2005: 22)

> Jezična sposobnost uključuje tradicionalne gramatičke razine.
> 1. *Leksička sposobnost* uključuje znanje i mogućnost uporabe rječnika, tj. leksičke riječi (višečlane riječi, frazemi, idiomi, kolokacije...) i gramatičke riječi (članovi, pokazne riječi, osobne zamjenice, upitne, odnosne i posvojne zamjenice, prijedlozi, pomoćni glagoli, veznici, čestice...).
> 2. *Gramatička sposobnost* uključuje obilježja pojedinih elemenata kao što su vrste morfema, riječi, kategorije (rod, broj, padež, vid, konkretno i apstraktno, brojivo i nebrojivo, prijelazno i neprijelazno...), vrste (sprezanje, sklanjanje, vrste riječi), strukture (izvedenice i složenice, sintagme,

surečenice...), procese (nominalizacija, stupnjevanje, preoblike...) i odnose (upravljanje, valentnost...).

3. *Semantička sposobnost* omogućuje ustroj značenja: leksičkoga (značenja riječi), gramatičkoga (značenja gramatičkih elemenata, kategorija, struktura i procesa) te pragmatičkoga (logički odnosi kao što su pretpostavke itd.).
4. *Fonološka sposobnost* uključuje vještine opažanja i proizvodnje glasova (fonema i alofona, razlikovnih obilježja), slogova, naglasnih obilježja, prozodijskih obilježja...
5. *Ortografska sposobnost* uključuje tiskano i rukopisno pismo, velika i mala slova, pravopisni zapis riječi, kratice, dijakritičke znakove, tipografske inačice, opće logografske znakove itd.
6. *Ortoepska sposobnost* uključuje čitanje ili govor koji se primio samo u pisanome obliku, mogućnost razrješavanja homonimskih ili rečeničnih dvoznačnosti u kontekstu... (Zanimljivo je da se tumačenje ove sposobnosti razlikuje od uobičajenoga hrvatskoga, jer ortoepija na hrvatskome uključuje standardan izgovor jezika, ne samo pisanoga jezika.) (Jelaska 2005a: 22)

Razlozi ograničenja na samo dva elementa jezične kompetencije (stavke leksičke i gramatičke sposobnosti) kao polazišta za analizu leže u činjenici da se ishodima ovoga rada podrazumijeva izrada materijala koji neizravno pridonose njihovomu razvijanju.

Osim toga, važnost je razvijanja gramatičke sposobnosti pri učenju i poučavanju hrvatskoga kao drugoga ili stranoga jezika navedena i u *Kurikulumu hrvatske nastave u inozemstvu*[14], gdje se među ostalim navodi:

> Nastava hrvatskoga jezika temelji se na cjelovitim tekstovima ili govornim situacijama: glas, riječ, sintagma, rečenica ne izdvajaju se od tekstovne situacije (konteksta). Nastava polazi od tekstova (lingvometodički predlošci), a učenik mora usvajati hrvatski jezik na raznovrsnim sadržajima, tekstovima i funkcionalnim stilovima (razumijevanje i interpretacija tekstova). Na temelju načela analize i sinteze tekstovi se raščlanjuju na strukturne sastavnice, utvrđuje se značenje i funkcija jezičnih činjenica, provodi se generalizacija i uspoređuju stečene spoznaje. Jezične djelatnosti na hrvatskom standardnom jeziku: slušanje, govorenje, čitanje, pisanje, ostvaruju se na sadržajima koji su aktualni, učenicima bliski, zanimljivi, poznati, korisni i obuhvaćaju aktualna jezična pitanja. Nastava hrvatskoga jezika pretpostavlja **istraživanje jezičnih pojava i zakonitosti** i postupno uvođenje učenika u jezično stvaralaštvo. Učenik je istraživač i pred njega se postavlja određeni **jezični problem**, pitanje, teškoća, dvojba. **Gramatičko mišljenje razvija se promatranjem i samostalnom analizom jezičnih činjenica, otkrivanjem njihovih osobina, rješavanjem gramatičkih zadaća, ispravljanjem jezičnih pogrješaka.** Provedbom školskih jezičnih pokusa učenici se usmjeravaju na jezične probleme – jezično zamišljanje i mišljenje (pokusi zvučnosti, premještanja

[14] *Kurikulum hrvatske nastave u inozemstvu* preuzet je s internetskih stranica Ministarstva prosvjete, kulture i športa Republike Hrvatske, http://public.mzos.hr/Default.aspx?sec=2116, 6. lipnja 2008; dalje u tekstu: Kurikulum

> zamjene, preoblike; stvaranje). Povezivanje jezične teorije i prakse ostvaruje se stvaranjem tekstovnih jedinica, sastavljanjem rečenica, iskaza, tekstova prema uzorku – **stvaranje tekstova prema zadanom gramatičkom uzorku**[15] (http://public.mzos.hr/Default.aspx?sec=2116, 4. lipnja 2008: 90-91).

Kao što je iz navedenoga citata vidljivo, gramatička sposobnost zauzima prilično važno mjesto pri učenju i poučavanju hrvatskoga kao drugoga ili stranoga jezika. Velika bi se pozornost pri poučavanju hrvatskoga jezika trebala posvetiti gramatičkim zadatcima receptivnoga (*promatranje*) i produktivnoga tipa (*stvaranje tekstova prema zadanom gramatičkom uzorku, ispravljanje pogrješaka*). Isto tako u nastavi se favorizira induktivni pristup usvajanja gramatike (*učenik je istraživač, otkrivanje gramatičkih osobina, istraživanje jezičnih pojava i zakonitosti*).

Uzimajući u obzir odrednice koje su definirane *Kurikulumom,* analizom nastavnih materijala, a u svrhu daljnje usporedbe dobivenih rezultata s navodima u postojećim gramatičkim priručnicima te potom rješenjima koja će biti primijenjena pri sastavljanju morfološkoga rječnika, odgovorit će se na sljedeća pitanja:

1. Koje su riječi i vrste riječi u najvećoj mjeri zastupljene u analiziranim materijalima te na kojem se mjestu one nalaze s obzirom na čestotnost njihova pojavljivanja u hrvatskome jezičnom korpusu?
2. Koje su gramatičke i leksičke kolokacije (bigrami i trigrami) najzastupljenije u nastavnim materijalima te na kojem se mjestu one nalaze s obzirom na učestalost i vjerojatnost njihova pojavljivanja u referentnome korpusu?

Analizom postojećih gramatika i gramatičkih priručnika, uzimajući u obzir postavljena težišta i rezultate, odgovorit će se na sljedeća pitanja:

1. U kojoj mjeri prikazi i objašnjenja morfoloških karakteristika i/ili paradigmi najčešće korištenih riječi i sastavnica kolokacija pridonose njihovome usvajanju kod učenika hrvatskoga kao drugoga i stranoga jezika?
2. Pružaju li gramatički priručnici učenicima hrvatskoga kao drugoga ili stranoga jezika konkretne informacije o paradigmama riječi koje su se pokazale kao najčešće pojavnice u nastavnim materijalima?

[15] Isticanja moja

Kako bi se dobili što objektivniji odgovori na postavljena pitanja pri analizi materijala korišten je proces korpusne analize. Razlozi odabiranja procesa korpusne analize leže u njezinoj objektivnosti i mogućnosti potvrde dobivenih rezultata. Prema riječima Lemniztera i Zinsmeister (2006: 10) pridržavanje određenih načela osnovni je preduvjet svake znanstvene djelatnosti, a tomu pripada i činjenica da se rezultati istraživanja moraju moći provjeriti ili čak po potrebi reproducirati[16]. Kao testni korpus korišteni su sljedeći nastavni materijali (Drilo, Stjepan (1996). *Kroatisch Tl.2. Lehrbuch für Fortgeshrittene* i Drilo, Stjepan (2006). *Kroatisch Tl.1. Lehrbuch für Anfänger*), dok je kao referentni korpus korišten korpus koji sam sama sastavila (više o tome u poglavlju 6.3.). Navodi o učestalosti pojavljivanja leksema u hrvatskome jezičnome korpusu citirani su iz *Hrvatskog čestotnog rječnika*[17].

2.1.1. *Korpusna analiza*

Postupak analize nastavnih materijala sastojao se od procesa konvertiranja materijala u oblik tekstualne datoteke, formata koji podržavaju računalni programi za statističku analizu. Važno je pri tome napomenuti da se pri konvertiranju materijala u tekstualnu datoteku pozornost obratila posebice na pročišćavanje teksta, odnosno brisanje svih sastavnica koje se sa stajališta ciljeva korpusne analize smatraju nebitnima (riječi i rečenica iz stranih jezika, zaglavlja i broja stranica).

Za dobivanje podataka o učestalosti pojavljivanja riječi korišten je računalni program *dtsearch*[18], dok je za analizu kolokacija korišten program *Ngram Statistics Package*[19]. U nastavku slijedi podrobniji opis načela rada navedenih programa.

[16] Prijevod MA. Izvornik: „Die Einhaltung gewisser Prinzipien ist die Grundvoraussetzung jeder wissenschaftlichen Tätigkeit. Dazu gehört, dass die Ergebnisse von Unersuchungen nachprüfbar oder sogar reproduzierbar sein müssen."
[17] Dalje u tekstu: HČR (Moguš-Bratanić-Tadić 1999)
[18] http://www.dtsearch.com/customDevelop.html, 29. travnja 2008.
[19] http://www.d.umn.edu/~tpederse/nsp.html, 29. travnja 2008.

2.1.2. *Dtsearch*

Da bi se dobio uvid u učestalost korištenja riječi i izraza u nastavnim materijalima za učenje hrvatskoga kao drugoga ili stranoga jezika, moralo se odlučiti na automatsko prebrojavanje riječi dobivenoga korpusa. Za tu se svrhu najpogodnijim učinio program *dtsearch*, koji je razvila skupina stručnjaka s ciljem brzoga pretraživanja velikoga broja tekstova.

Dtsearch naime omogućuje indeksiranje svih leksema iz nekoga teksta te na taj način statistički obrađuje zadani korpus. Važno je pri tome napomenuti da sam program bilježi riječi u svome zadanome obliku te da ih ne lematizira, što otežava daljnju statističku obradu. Za lematizaciju dobivenih rezultata koristila sam se lematizatorom koji sam razvila u suradnji s tvrtkom MorphoLogic.

Namjena softvera *dtsearch* u ovome istraživanju bila je utvrđivanje najčešćih pojavnica koje se rabe u nastavnim materijalima za učenje hrvatskoga jezika na višim razinama te uspoređivanju dobivenih rezultata s navodima u HČR-u i testnome korpusu. Uspoređivat će se naime najčešće pojavnice – leme i oblici riječi. Nadalje, na osnovi dobivenih rezultata bit će provedena i analiza jezičnih priručnika te će se utvrditi prioriteti pri izradi morfološkoga rječnika hrvatskoga jezika.

2.1.3. *NSP*

Za određivanje daljnjih prioriteta pri izradi morfološkoga rječnika hrvatskoga jezika potrebno je utvrditi i tipove kongruencije koji se pojavljuju u korpusu te odrediti najčešće pojavnice. U tu je svrhu korišten slobodno dostupan softver NSP, čija je namjena kao i opis rada navedena u odlomcima koji slijede.

NSP, ili N–gram Statistic Package besplatan je i slobodno dostupan program na temelju programskoga jezika Perl, koji je razvila skupina stručnjaka na čelu s Tedom Pedersenom. Program je nastao kao proizvod zajedničke suradnje matematičara i lingvista, a temelji se na načelima matematičke i statističke analize. Osim mnogih znanstvenih područja na kojima se može primijeniti, program se uspješno primjenjuje i za istraživanja na području jezikoslovlja.

NSP prije svega služi svrsi korpusne analize. Kada je riječ o korpusima, potrebno je napomenuti da NSP omogućuje njihovu brzu i ciljanu pretragu u svrhu traženja kolokacija[20], odnosno relacija između dviju, tri, četiri te teoretski neograničenoga broja riječi, što se može smatrati i primarnim područjem rada samoga programa. Relacije između riječi ne ispituju se samo u neposrednome području, nego program analizira i odnose ili poveznice između primjerice dviju riječi, čak i ako se one ne nalaze u neposrednoj blizini, odnosno ako ih dijeli nekoliko jezičnih elemenata – riječi ili rečeničnih znakova. Sekundarnim područjem djelovanja NSP-a može se smatrati brza i ciljana pretraga unutar dobivenih rezultata u svrhu traženja relacija ili odnosa jedne kolokacije prema ostalim jezičnim elementima, što se može iskoristiti i za daljnju pretragu i istraživanje drugih frazeoloških jedinica, primjerice idioma.

Način rada i pretraživanja korpusa uz pomoć NSP-a događa se u nekoliko koraka. Prva se faza sastoji od pripremanja tekstova za analizu, što se događa u dva koraka. Prvi naime podrazumijeva pročišćavanje teksta, odnosno brisanje sastavnica i onih znakova koji se sa stajališta ciljeva opisanoga istraživanja smatraju nepotrebnima. Iz razloga što se kao cilj istraživanja postavilo automatsko traženje frazeoloških jedinica, odnosno kolokacija, pod nepotrebnim se znakovima podrazumijevaju interpunkcijski znakovi te slučajne pogrješke u smislu upotrebe nestandardnih znakova. Iako sam program omogućuje automatsko brisanje nepotrebnih znakova u tekstu, ne preporučuje se korištenje navedene opcije iz razloga što se tako dobiveni rezultati ne mogu smatrati potpuno točnima. Problem se, naime, kod takve automatske obrade znakova uz pomoć NSP-a pokazuje u brisanju primjerice svih zareza, crtica, povlaka i zagrada, odnosno problem se pojavljuje ako se neki od znakova pojavljuju kao dijelovi riječi. Automatskim se brisanjem znakova naime dobivaju nevažeći podatci, odnosno pogrješno napisane leme koje nisu u sastavu leksikona te koje se kasnije ne mogu pravilno analizirati. Kao primjer jedne takve riječi može se navesti *m/b*, skraćenica od riječi *motorni brod*, koja se pojavljuje i u tekstovima na hrvatskome jeziku, a kod koje je vidljivo da kosa crta čini sastavnicu riječi. Korištenjem opcije automatskoga brisanja skraćenica se pretvara u *m b,* postaje nevažeća i svrstava se u

[20] Pojednostavljena definicija izraza *kolokacija* u ovome radu podrazumijeva učestalo pojavljivanje dvaju ili više jezičnih jedinica, gdje je vjerojatnost da će se te jedinice uvijek pojavljivati zajedno vrlo velika. Ona proizlazi iz Bensonsove definicije (1990, opus.cit. http://www.latl.unige.ch/personal/vseretan/publ/EURALEX2004_VS_LN_EW.pdf, 20. siječnja 2008.) da su kolokacije „proizvoljne kombinacije riječi koje se ponavljaju“ te stajališta Lemnitzera i Zinsmeister (2006:15) da se izraz *kolokacija* odnosi na *kookurenciju* (*Kookkurenz, Kovorkommen*). U ovome se radu izraz *kolokacija* koristi za obilježavanje i gramatičkih kolokacija, odnosno kookurencija.

pogrješku. Posebnim se pročišćavanjem teksta prije upotrebe samoga računalnog programa NSP izbjegavaju navedene pogrješke te se time omogućuje poluautomatsko brisanje svih onih znakova koji se ne smatraju dijelovima hrvatskoga leksikona.

Drugi važan korak pri pripremi tekstova za analizu uz pomoć programa jest lematizacija, odnosno analiza tekstova na morfološkoj razini te pridodavanje svih oblika zadane riječi odgovarajućoj lemi, odnosno osnovi riječi. Lematizacija hrvatskih tekstova nešto je teži postupak jer je NSP izrađen prije svega za analizu tekstova na engleskome jeziku, koji se s morfološkoga gledišta ubraja u jezike smanjene flektivnosti (u usporedbi s hrvatskim ili nekim drugim slavenskim jezikom). Na tržištu je dostupno nekoliko besplatnih lematizatora, međutim iz razloga što njihovi dometi nisu odgovarali postavljenim zahtjevima istraživanja, pri radu s NSP-om koristila sam se svojim računalnim morfološkim analizatorom hrvatskoga standardnog jezika. Razlozi lematizacije leže u činjenici da je neophodno imati alat koji će sve oblike jedne riječi povezati s lemom, odnosno neće ih tretirati kao zasebne leksičke jedinice. Primjerice, u djelomično idiomatiziranome izrazu *mlatiti praznu slamu* glagol *mlatiti* naime može se pojaviti u različitim glagolskim vremenima i licima. Vrlo se važnim korakom smatra proces lematizacije kojim se primjerice oblici poput *mlatio* u obliku ***mlatio je*** *praznu slamu* ili *mlatiše* u obliku ***mlatiše*** *praznu slamu* povežu sa svojom lemom – infinitivnoj osnovi glagola *mlatiti* te da se sukladno tome analiziraju. Rezultati dobiveni na taj način smatraju se točnijima jer izravno utječu na točne izračune korelativnih odnosa među sastavnicama. Isto je tako važno napomenuti da se tijekom analize odnosno lematizacije tekstova zanemaruju razlike između velikih i malih slova u riječima. Kod daljnje analize rezultata uzimaju se u obzir velika i mala slova, što je važno posebice ako ona imaju razlikovnu funkciju. Primjerice, na potpovršinskoj se razini za vrijeme morfološke analize uzima u obzir razlikovna uloga velikih i malih slova te se riječi poput *pleše* (3. lice jednine glagola *plesati*) i *Pleše* (vlastito ime – *Vesna Pleše*) lematiziraju na odgovarajući način, odnosno ne povezuju se s istom lemom.

Prije početka pretrage korpusa, a nakon njegova sastavljanja, važnim se korakom pokazalo definiranje najmanjih pretraživačkih jedinica, takozvanih *tokena*. Uzimajući u obzir ciljeve postavljenoga istraživanja, tokenom se u užem smislu definiraju riječi, odnosno nizovi znakova koji se nalaze u tekstu između dvaju praznih znakova (razmaknica) ili interpunkcijskih znakova. Važno je napomenuti da razmaknica odnosno prazan znak u ovome programu ne predstavlja granice obrade i analize. Drugim riječima, tokeni se u

programu NSP mogu slobodno definirati, što znači da ako se provodi sekundarno istraživanje u smislu ciljane pretrage kolokacija određene skupine riječi, token se može sastojati od cijele fraze, odnosno skupine od dviju, tri ili teoretski neograničenoga broja riječi. Time se podrazumijeva da se kao token može uzeti primjerice ne samo imenica *golub*, nego i imenske fraze poput *bijeli golub* ili *golub mira.* Nakon definiranja takve vrste tokena program pretražuje poveznice između riječi ili zadane skupine riječi kao tokena i ostalih riječi u korpusu.

Nadalje, definiranjem tokena određuju se daljnja područja analize te onemogućuju dvosmisleni rezultati odnosno dobivanje nevažećih rezultata kao kolokacija. Primjerice, ako se tokenom postavi imenica *ruža*, program će izbaciti različite rezultate od onih ako se tokenom postavi jedinica *Ruža Pospiš Baldani,* čime se mijenjaju same interpretacije dobivenih rezultata. Osim navedenih razloga, definiranje se tokena pokazalo iznimno važnim kod analize rezultata u pogledu relacija između lijevo odnosno desno pozicioniranih tokena ili jedinica pretraživanja. Primjerice, ako se promatra riječ *jabuka*, velika je vjerojatnost da će se ona često pojavljivati u kolokacijama s pridjevom *crven* čineći tako imensku frazu *crvena jabuka*. Ako se, međutim, promatra samo lijevopozicionirani token *crven*, nije vjerojatno da će ta riječ tvoriti najviše kolokacija s imenicom *jabuka* ili da će se u većini slučajeva ta imenica nalaziti u neposrednoj blizini kao njezin najvjerojantniji desnopozicionirani token[21].

Nakon analize tekstova, kao sljedeći modul programa NSP-a slijedi obrada dobivenih rezultata uz pomoć različitih matematičkih i statističkih operacija. Rezultati se broje (count.pl) i statistički obrađuju (statistic.pl). Njihovo se ocjenjivanje i rangiranje nakon toga obavlja uslijed određivanja vjerojatnosti njihova pojavljivanja u cjelokupnom jezičnom opusu, što se pokazalo iznimno korisnim. Izračunom vjerojatnosti i učestalosti pojavljivanja kolokacije izbjegavaju se nepouzdane informacije. Iako se možda jedna kolokacija pojavljuje često u danome korpusu, ona se ne može ocijeniti važećom ako su njezine sastavnice zastupljene u samo malom broju u tekstovima. Program izračunava, primjerice, da su dvije sastavnice velikom vjerojatnošću uvijek u danom korelativnom odnosu, ali zbog rijetkosti pojavljivanja zadanih lema u cjelokupnome korpusu korelacija ne može se smatrati važećom u odnosu na cijeli jezični opus. Statistička se obrada nadalje

[21] Pri korpusnoj analizi referentnoga korpusa kao rezultati istraživanja lijevopozicioniranih tokena imenice jabuka najčešće su se pojavili pridjevi *zlatna* i *rumena*, dok se kao desnopozicionirani token pridjeva *crven* najčešće pojavljuje imenica *križ* u kolokaciji *crveni križ*.

pokazala vrlo korisnom jer omogućuje interpretaciju pretrage unutar već dobivenih rezultata te njihovo rangiranje prema stupnju valentnosti, od gore prema dolje – od visokoga do najnižega stupnja valentnosti.

U sljedećem će se odlomku podastrijeti rezultati analize nastavnih materijala za učenje hrvatskoga jezika na njemačkome govornom području uz pomoć ovih dvaju programa.

2.2. Analiza nastavnih materijala za učenje hrvatskoga jezika na njemačkome govornom području

Analiza materijala koji se rabe na njemačkome govornom području određena je ciljevima ovoga rada, odnosno izradom morfološkoga rječnika hrvatskoga jezika za govornike njemačkoga jezika. Cilj je naime ove analize pomoću odgovora na pitanja postavljenih u poglavlju 2.1. utvrditi vrstu leksika koja je zastupljena u udžbenicima s obzirom na odrednice temeljnoga rječnika, te pokrivaju li jezični priručnici koji se rabe u nastavi hrvatskoga kao drugoga ili stranoga jezika jezične sadržaje koji su predodređeni udžbenicima. Iako su slične analize udžbenika za učenje hrvatskoga kao drugoga ili stranoga jezika već provedene (vidi Blagus 2005a i 2005b), s obzirom na ciljeve ovoga istraživanja opisanih u prethodnim odlomcima analiza će se nastavnih materijala temeljiti na kvantitativnoj analizi udžbeničkoga korpusa te će se za potrebe morfološke analize i izrade računalnoga morfološkog rječnika hrvatskoga jezika usredotočiti na područje zastupljenosti leksika prema vrstama riječi u udžbenicima, čestoći pojavnica i korelativnim odnosima među riječima. Blagus (2005a) je naime u svome radu provela istraživanje ciljanoga leksika, odnosno svoja saznanja potkrepljuje navođenjem primjera jednoga tematskoga područja – hrane i pića. Usredotočujući se na udžbenike za odrasle koji se rabe prije svega za govornike engleskoga jezika odnosno hrvatska izdanja koja se rabe na nastavi hrvatskoga jezika u inozemstvu, Blagus je postavila temelj za daljnja istraživanja u tome području te pridodaje da bi se „daljnjim leksičkim analizama mogla dobiti cjelovitija udžbenička građa kojom bi se u budućnosti olakšalo sastavljanje udžbenika (Blagus 2005a: 273).“ U ovome dijelu rada provest će se ciljano istraživanje onih udžbenika koji se rabe na njemačkome govornom području za učenje i poučavanje hrvatskoga kao drugoga ili stranoga jezika, a koji nisu pokriveni istraživanjima spomenute autorice.

Pri izboru samih nastavnih materijala koji su ušli u izbor, postavilo se nekoliko kriterija. Za sastavljanje udžbeničkoga korpusa na kojem će se temeljiti ovo istraživanje, kao referentni izvor rabio se *Kurikulum hrvatske nastave u inozemstvu*[22], točka 8 te rad Lidije Cvikić (2005b) koja je dala pregled udžbenika za govornike njemačkoga jezika (2005b: 221). Uspoređivanjem već analizirane udžbeničke građe iz radova Blagus (2005a i 2005b) te Cvikić (2005b) zaključilo se kako među materijale koji do sada nisu podvrgnuti računalnoj kvantitativnoj analizi pripadaju udžbenici S. Driloa *Kroatisch, TL 1, Lehrbuch für Anfänger* (2006)[23] i *Kroatisch Lehrbuch für Fortgeschrittene* (1996)[24]. Iz razloga što bi se računalnim morfološkim rječnikom hrvatskoga jezika mogli koristiti svi govornici, odlučilo se, radi usporedbe rezultata, za analizu leksika ovih dvaju izdanja. Rezultati se ove analize iz toga razloga neće smatrati rezultatima referentnoga udžbeničkog korpusa nastavnih materijala za strance, nego će samo pružiti uvid u građu kojom se koristi na njemačkome govornom području te može poslužiti za opsežnija istraživanja kako bi se postigli navedeni ciljevi.

Analizi navedenih udžbenika pristupilo se sa stajališta korpusne lingvistike. Tekst je udžbenika pročišćen od suvišnih elemenata, poput crtica, brojeva stranica i slično, te je podijeljen s obzirom na različita gledišta analize. U analizu su ušla samo ona područja koja su pisana hrvatskim jezikom. Njemačka objašnjenja i gramatički dio izostavljeni su. Korpus je lematiziran, lematizacija je obavljena uz pomoć alata koji je napravljen na temelju morfološkoga analizatora HUMOR-a. Iz razloga što još ne postoji disambiguator za hrvatski jezik (alat koji bi, pojednostavljenim riječima, sam procijenio koji od više jednakih oblika odgovara određenoj lemi), dvojbeni su oblici lematizirani ručno. Rezultati će se analize predstaviti za svaki udžbenik pojedinačno i skupno, a cjelokupni će se materijal usporediti s obzirom na temeljni rječnik njemačkoga jezika.

Ako se pogledaju naslovi navedenih udžbenika, zaključuje se da udžbenici pokrivaju leksik od početne do srednje razine. Iz predgovora navedenih udžbenika postaje vidljivo da se radi o prerađenim izdanjima i da ona obuhvaćaju područje hrvatskoga standardnog jezika. Teme koje su primjerice pokrivene udžbenikom KF zadiru čak u područje dijalektologije odnosno inačice hrvatskoga jezika. Osim natuknica iz hrvatskoga standardnoga jezika, KF

[22] *Kurikulum hrvatske nastave u inozemstvu,* točka 8 nosi naziv *Literatura i izvori za pripremu i izvođenje nastave*. http://public.mzos.hr/Default.aspx?sec=2116-a, 10. lipnja 2008.
[23] Dalje u tekstu KA
[24] Dalje u tekstu KF

sadrži i leme iz čakavskoga i kajkavskoga narječja te opis oblikotvornih i fleksijskih morfema koji se rabe za tvorbu riječi i oblika u tim narječjima. Na prvi se pogled time očekuje da su tematski pokrivena područja koja su definirana i temeljnim rječnikom. Iz razloga što se temeljni rječnik za hrvatski jezik još nije utvrdio (Jelaska-Cvikić-Novak-Milić 2005: 197-198), može se samo pretpostaviti da on podrazumijeva 2000 odnosno 4000 najčešćih riječi. Jelaska-Cvikić-Novak-Milić (2005) tako navode:

> Od stranih izvora korisni su prijevodi temeljnih rječnika za neke druge jezike, kao što su npr. *Grundwortschatz Enlgisch* (Langenscheidt 2000) i *Basic German Vocabulary* (Langenscheidt 1998) jer su njima obuhvaćene teme i pojmovi zajednički svim ljudima, iako ne i sve riječi. (…) Usporedbom građe iz takvih izvora vjerojatno bi se dobio prikladan popis najčešće dvije tisuće, odnosno četiri tisuće riječi (Jelaska–Cvikić–Novak-Milić 2005: 199).

U udžbenicima KA i KF ukupno postoji 7591 različnica. Lematizacijom je utvrđeno da se radi o 3833 natuknice[25]. Sve su različnice svedene na mala slova da bi se olakšao postupak lematizacije, jer lematizator razlikuje mala i velika slova. U ovom će se poglavlju, zbog analize različitih oblika, predočiti nelematizirani primjeri.

Ako se tematski prouče natuknice koje se nalaze u spomenutim izdanjima te kada se one usporede s Langenscheidtovim (1998) izdanjem *Basic German Vocabulary*, uočava se da postoje neke natuknice koje ne pripadaju popisu od 4000 najčešćih riječi koje se pojavljuju u njemačkome jeziku (a za koje se pretpostavlja da bi se analogno trebale pojaviti u sastavu temeljnoga rječnika hrvatskoga jezika), primjerice *bezutješan, božanski, brkati* (glagol), *diletantizam, kolebanje* itd[26]. Time se dolazi do zaključka da se u udžbenicima nalazi leksik koji po svemu sudeći prelazi sastavnice temeljnoga rječnika. Iz razloga što morfološki rječnik *Morko* po svemu sudeći sadrži samo temeljnu rječničku građu[27], zaključuje se, ako bi se htjelo postići da se odgovori zahtjevima učenika hrvatskoga jezika i na višim razinama, postoji potreba za izradom morfološkoga rječnika hrvatskoga

[25] Važno je pri tome napomenuti da su se glagolski prilog prošli i sadašnji svrstali u kategoriju priloga, a ne u glagolsku paradigmu s glagolom kao natuknicom (kao što je on tretiran u HUMOR-u). Isto su se tako tretirali i prilozi koji su nastali od pridjeva. Iako oni imaju isti oblik kao i neki oblici pridjeva, ako se ručnom lematizacijom utvrdilo da se radi o prilogu, oni su uvršteni pod zasebnu natuknicu.

[26] Uzimajući u obzir postavljene ciljeve ovoga rada i istraživanje udžbenika, nije se detaljnije analizirao njemački temeljni rječnik, niti prevodile natuknice na hrvatski jezik, nego su primjeri navedeni samo za egzemplifikaciju problema. U jednome od budućih projekata prevodit će se natuknice iz njemačkoga temeljnoga rječnika i uspoređivati s hrvatskom građom.

[27] Morfološki rječnik prema mojim informacijama nije dostupan u široj prodaji te nisam bila u mogućnosti pogledati ga. Informacije o *Morku* odnosno *Morfeku* preuzete su s internetske stranice http://www.croatiana.org/croatiana--projekti-hrvatskikaodrugi.htm, 28. lipnja 2008. godine

standardnog jezika koji prelazi razine temeljnoga rječnika i pokriva leksik koji se nalazi u primjerice Anićevu rječniku.

Analiza različnica prema vrstama riječi u udžbenicima KA i KF dana je u sljedećem prikazu.

	Vrsta riječi	Broj različnica u KA	Broj različnica u KF	Ukupno prema vrsti riječi	Udio vrste riječi u tekstu
A	imenice	697	2855	3552	46,8%
B	glagoli	361	1388	1700	22,4%
C	pridjevi	232	1400	1632	21,5%
D	prilozi	22	215	237	3,1%
E	zamjenice	10	174	184	2,5%
F	prijedlozi	17	13	30	0,4%
G	ostale vrste riječi	-	-	256	3,4%
∑	ukupno (A+B+C+D+E+F+G)	1339	6045	7591	

Prikaz 1: Broj različnica prema vrstama riječi u udžbenicima KA i KF

Kao što je iz priloženoga prikaza uočljivo, ukupan zbroj različnica prisutnih u udžbenicima KA i KF nije jednak zbroju različnica koje se pojavljuju u pojedinim udžbenicima. Razlog tomu jest pojavljivanje određenog broja riječi u oba udžbenika.
Iz priložene se tablice vidi da se broj hrvatskih riječi u drugome izdanju povećao za gotovo pet puta (4,51). To se percipira normalnim prirastom jer izdanje za početnike obiluje njemačkim popratnim tekstom i njemačkim objašnjenjima, dok je drugo izdanje bogatije hrvatskim leksemima. Ako se pretpostavi da je u izdanjima na intuitivnoj osnovi izabran temeljni leksik (Cvikić 2005b:221), te da se približno 2000 najčešćih pojavnica smatra temeljnim rječnikom, može se zaključiti da izdanje za naprednije učenike prelazi okvire temeljnoga rječnika što još jednom govori u prilog potrebi za izdavanjem opsežnijih izdanja kojima bi se koristili učenici hrvatskoga kao drugoga ili stranoga jezika.

Da bi se utvrdili prioriteti pri izradi morfološkoga rječnika, potrebno je pogledati čestoću pojavljivanja različnica u udžbenicima. Sljedeća tablica prikazuje izbor od trideset najčešćih različnica koje se pojavljuju u oba udžbenika prema vrstama riječi. Iz razloga što je cilj ovoga rada izrada morfološkoga rječnika hrvatskoga jezika s prikazima paradigmi,

za ovaj su prikaz izabrane promjenjive vrste riječi s najvećim udjelom pojavljivanja u udžbenicima KA i KF – imenice, pridjevi i glagoli.

Imenice		Pridjevi		Glagoli		Riječi ukupno	
Različnica	Broj pojavlj.	Različnica	Broj pojavlj.	Različnica	Broj pojavlj.	Različnica	Broj pojavlj.
ivan	132	braun	22	je	1909	je	1909
günther	127	studentski	10	se	992	i	1139
ana	115	zatvorena	8	su	713	u	1053
helga	112	zelena	8	sam	322	se	992
student	94	veliko	7	nije	274	su	713
kuće	91	tramvajsko	6	će	155	da	467
studenti	76	zaposlena	6	nisam	112	ne	420
marija	69	sestrin	5	nisu	106	na	409
knjigu	64	zelene	5	bio	102	to	398
dan	62	dozvoljena	4	bilo	89	li	336
godine	54	jako	4	bi	87	sam	322
seminaru	52	pješački	4	rekao	83	nije	274
knjiga	51	turistički	4	bila	71	što	267
dana	50	zaposlilo	4	bih	65	a	264
ljubica	50	zatvoreno	4	bili	62	od	255
petar	48	bijel	3	ima	60	o	217
studentica	46	bolesni	3	doći	58	s	208
put	43	bratova	3	vidio	52	za	188
profesor	42	brzi	3	biti	50	ja	172
gospodin	41	brzim	3	razgovarali	46	smo	165
more	39	glavna	3	nismo	45	kada	161
automobil	38	gornji	3	došli	44	će	155
jezik	38	gradski	3	jeste	44	kod	154
roman	38	lijepe	3	govori	41	mi	141
vremena	37	malena	3	može	40	on	140
stan	36	prodano	3	mogu	39	ivan	132
zagrebu	36	putnički	3	ćemo	38	günther	127
brata	34	siguran	3	bismo	33	gdje	127
mjesta	33	spavaću	3	jesi	33	još	124
jezika	32	svježe	3	odgovorio	33	ti	123

Prikaz 2. Popis trideset najčešćih različnica prema vrstama riječi u udžbenicima KA i KF

Proučavanjem navedenih različnica vidimo da podatci odgovaraju podatcima iz HČR. Iz ukupnoga broja pojavljivanja lema, glagol *biti* i u čestotnom se rječniku i u analiziranim udžbenicima pojavljuje na prvome mjestu. Podjednako su zastupljene imenice koje označuju živo i imenice koje označuju neživo, vlastita imena nisu izbačena s popisa jer imaju imeničku deklinaciju (više o tome u poglavlju 7.2.). Vidljivo je i iz rezultata analize

nelematiziranoga teksta da se neke imenice pojavljuju u više oblika, što znači da su one u cjelokupnom korpusu zastupljenije (primjerice *dan* i *jezik*).
S obzirom na ukupan broj riječi u tekstu pridjevi se ne pojavljuju vrlo često, odnosno u udžbenicima se ne ponavljaju pridjevski oblici. Vidi se da se od trideset najčešćih pridjeva petnaest pojavljuje svega triput u oba udžbenika. Iz razloga što se kod pridjeva radi o 1632 različnice, zaključuje se da udžbenici obiluju raznolikošću pridjeva i pridjevskih oblika.
Ako podrobnije analiziramo glagole koji se pojavljuju u udžbenicima, uočavamo da veliki broj pojavljivanja imaju nepravilni glagoli i pomoćni glagoli (*biti, htjeti)*, što odgovara i rezultatima iz HČR. Prema analizi različnica postoji osam oblika glagola u imperfektnom ili aoristnom obliku. Kao što se vidi, u ukupnom broju najčešćih različnica koje se pojavljuju u udžbenicima od promjenjivih su vrsta riječi na najvišem mjestu imenice, a zatim glagoli.

Rezultati ove analize dokazuju da ako se prioriteti pri izradi morfološkoga analizatora ili morfološkoga rječnika postave u skladu s rezultatima čestotne analize (i HČR-a, i za potrebe ovoga rada sastavljenoga udžbeničkog korpusa), povećava se prepoznatljivost nekoga teksta već uslijed izrade i testiranja programa. Što se tiče odrednica morfološkoga rječnika, postoje naznake da će učenike kod samih paradigmi ponajprije zanimati oblici onih lema koje su navedene u gornjem prikazu. Manja je vjerojatnost da će učenici dati prednost onim lemama koje su slabije zastupljene u udžbeničkome korpusu, što daje dobre smjernice pri izradi rječnika.

Uzimajući u obzir jezične varijante različnica koje se pojavljuju u korpusu, možemo uočiti da se različnice s dijalektalnim varijantama pojavljuju samo u udžbeniku za naprednije učenike, KF-u. Na žalost, isto tako taj udžbenik sadrži i većinu različnica koje su prema Brodnjakovu (1991) rječniku srbizmi, odnosno više se rabe u srpskome jeziku. Popis dvojbenih različnica s Brodnjakovim pojašnjenjima dan je u sljedećem prikazu:

Natuknica	Vrjedovanje prema Brodnjakovu (1991) *Rječniku razlika između hrvatskoga i srpskoga jezika*
sediti	ima uz značenja koja su različita u srpskom i hrvatskom jeziku i značenje koje je u oba jezika istovjetno
dopadati	srpska riječ; riječ koja u srpskom ima jedno značenje, u hrvatskom drugo

konsekvencija	riječ pripada (i) hrvatskom jeziku, općeprihvatljiva je riječ
nerv	tuđica, u srpskom obična riječ, u hrvatskom se zamjenjuje
surevnjiv	srpska riječ, hrv. jednakovrijednica: *jalan*
sveštenik	srpska riječ, hrv. jednakovrijednica: *svećenik*
zvezda	srpska riječ, hrv. jednakovrijednica: *zvijezda*
akcenat	srpska riječ, hrv. jednakovrijednica: *akcent*
opšti	srpska riječ, hrv. jednakovrijednica: *opći*
univerzitet	riječ pripada (i) hrvatskom jeziku, ali je tipičnija za srpski nego za hrvatski književni jezik
tačan	srpska riječ, hrv. jednakovrijednica: *točan*
vazda	srpska riječ, hrv. jednakovrijednica: *mnogo*
apoteka	riječ pripada (i) hrvatskom jeziku, općeprihvatljiva je riječ, tipičnija za srpski nego za hrvatski književni jezik, zabilježena u djelima starijih hrvatskih pisaca

Prikaz 3: Popis dvojbenih natuknica iz udžbenika s Brodnjakovom kategorizacijom

Ova činjenica dovodi nas do problema izbora jezične inačice pri izradi morfološkoga analizatora i rječnika, a koja je opisana u poglavlju 6.2.

Rezultati analize bigrama prema *t-testu* (Prilog 3) pokazuju da kod kolokacija dominiraju glagoli i prijedlozi kao tokeni. Time se zaključuje da su potrebe za informacijama u jezičnim priručnicima u pogledu glagolske i imeničke paradigme iznimno visoke.
Rezultati analize prema *Pointwise Mutual Information*[28] analizi (Prilog 4) prikazuju veliku važnost pridjevske paradigme te veću vjerojatnost pojavljivanja kolokacija ovakve vrste u samoj jezičnoj proizvodnji.
Više informacija u pogledu sintaktičkih odnosa među riječima u udžbenicima KA i KF može se dobiti analizom trigrama i 4-grama uz pomoć NSP-a (Prilozi 5 i 6). Iz razloga što je cilj ovoga rada morfološka analiza i izrada morfološkoga rječnika, problemima koji zadiru u sintaksu neće se podrobnije baviti.

[28] Pointwise mutual information (PMI) statistički je izračun kojim se predočava vrijednost odnosa između vjerojatnosti pojavljivanja para ili skupine ako se uzme u obzir njihovo zajedništvo u odnosu na vjerojatnost njihova pojavljivanja ako se uzme u obzir njihova pojedinačna distribucija i neovisnost. http://en.wikipedia.org/wiki/Pointwise_mutual_information, 29. srpnja 2008.

Uzimajući u obzir postavljene ciljeve ovoga rada i usredotočujući se na morfologiju hrvatskoga jezika, vidimo da su analizom udžbeničkoga materijala dobivene korisne informacije što se tiču paradigmi koje bi za učenike bile od najveće važnosti. U sljedećem odlomku slijedi analiza jezičnih priručnika s naglaskom na opisu i prikazu onih paradigmi u gramatikama i jezičnim priručnicima za koje se ispostavilo da su najviše zastupljene u nastavnim materijalima za učenje hrvatskoga jezika na njemačkome govornom području.

2.3. Analiza jezičnih priručnika koji se rabe u nastavi hrvatskoga jezika[29]

Pri analizi jezičnih priručnika izabrane su gramatike i jezični priručnici koji se rabe u nastavi hrvatskoga kao drugoga ili stranoga jezika, izdanja koja je odobrilo Ministarstvo prosvjete i športa Republike Hrvatske[30]. Osim navedenih izdanja izbor je za analizu sastavljen na temelju ankete provedene među nastavnicima hrvatskoga jezika koji podučavaju učenike kojima hrvatski nije materinski jezik[31]. Konačan se izbor temeljio na učestalosti uporabe određenoga izdanja za potrebe učenja i podučavanja hrvatskoga kao stranoga jezika na razinama B1 i B2. Iz razloga što se težilo što reprezentativnijem uzorku materijala za analizu, poštovalo se i načelo različitosti pristupa jezičnome sadržaju te su zbog toga naposljetku korpus za analizu činila sljedeća tiskana izdanja: J. Silić–I. Pranjković (2005) *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*[32], E. Barić et al. (1995) *Hrvatska gramatika*[33], D. Raguž (1997) *Praktična hrvatska gramatika*[34] te elektroničko izdanje *Gramatičkoga tezaurusa hrvatskoga jezika v. 1.2.* autora J. Silića, S. Batanožića i B. Ranilovića (1996)[35]. Radi usporedbe određenih navoda testirano je i

[29] Pri spomenu nastave hrvatskoga jezika, kao što je već objašnjeno u samome uvodu ovoga rada, misli se na nastavu hrvatskoga kao drugoga ili stranoga jezika na razinama B1 i B2. U ovome radu pod pojmom *učenici,* ako nije drugačije navedeno, podrazumijevaju se učenici za koje se pretpostavlja da im je jezična kompetencija na razini B1 odnosno B2.

[30] U izbor su ušla samo ona izdanja koja se odnose na stavku 8 *Kurikuluma hrvatske nastave u inozemstvu,* pod točkom 8.1.3 – jezikoslovlje (http://public.mzos.hr/Default.aspx?sec=2116-a, 10. lipnja 2008). Ostala se literatura ne rabi izričito za poučavanje gramatike hrvatskoga kao stranog jezika na razinama B1 i B2 odraslim učenicima, te je zbog toga izostavljena iz analize.

[31] Zbog karakteristika ovoga rada i nemogućnosti kontaktiranja nastavnika koji podučavaju hrvatski jezik na njemačkome govornom području, interna je anketa provedena među nastavnicima koji podučavaju hrvatski jezik mađarskim govornicima, odnosno govornicima kojima hrvatski nije materinski jezik u Prosvjetno-kulturnom centru Mađara u Republici Hrvatskoj, u Osijeku. U anketi je sudjelovalo četiri profesora hrvatskoga jezika.

[32] Dalje u tekstu: GHJ

[33] Dalje u tekstu: HG

[34] Dalje u testu: PHG

[35] Dalje u tekstu: GT

internetsko izdanje *Hrvatskoga morfološkoga leksikona*[36]. Razlog uvrštavanja HML u analizirana izdanja leži u riječima autora koji navode: „Leksikon može biti od koristi kako učenicima hrvatskoga jezika (kako izvornim govornicima, tako i strancima koji uče hrvatski), tako i stručnjacima i sustavima za pretraživanje (internet i intranet tražilice), crpljenje obavijesti, dubinsku obradbu teksta i računalno-lingvističku obradbu hrvatskih tekstova[37]“. Važno je pri tome napomenuti da, iako primarna svrha ovoga leksikona nije poučavanje hrvatskoga kao drugoga ili stranoga jezika te da zbog toga leksikon ne navodi uz oblike i tradicionalne gramatičke kategorije, nego morfosintaktički opis (MSD) usklađen s MulTextEast v 3.0 preporukama za hrvatski jezik[38], leksikon pruža mnogo korisnih informacija koje se tiču paradigmi i zastupljenih oblika. HML prema riječima autora obrađuje „više od 45.000 riječi općega jezika, 15.000 osobnih muških i ženskih imena i 50.000 prezimena registriranih u Republici Hrvatskoj. Na temelju toga rječničkoga blaga proizvedeno je više od 3.900.000 njihovih oblika[39].“

Glavni je cilj analize navedenih materijala bio utvrditi u kojoj mjeri sadržaji predočeni u izdanjima pridonose točnome rješavanju zadataka u udžbenicima hrvatskoga kao drugoga ili stranoga jezika, odnosno pružaju li jasne i precizne odgovore na postavljena pitanja.
Iz razloga što je cilj ovoga rada izrada morfološkoga rječnika, analiza će navedenih materijala obuhvatiti samo područje morfologije promjenjivih vrsta riječi suvremenoga hrvatskog jezika sa stajališta učenika hrvatskoga kao drugoga ili stranoga jezika te sljedeća gledišta:

1. jasnoća objašnjenja,
2. uporabljivost objašnjenja.

Podrobnija analiza problematike koja se tiče morfološkoga opisa pojedinih vrsta riječi slijedi u poglavlju 7. Važno je pri tome napomenuti da iz razloga što je cilj ovoga rada izrada morfološkoga rječnika te alata i programa za prevođenje pisanih tekstova sa i na

[36] Dalje u testu: HML
[37] http://hml.ffzg.hr/hml/info.php?show=hml, 10. lipnja 2008.
[38] http://hml.ffzg.hr/hml/info.php?show=hml, 10. lipnja 2008.
[39] http://hml.ffzg.hr/hml/info.php?show=hml, 10. lipnja 2008.

hrvatski standardni jezik koji se temelji na računalnome morfološkom opisu hrvatskoga standardnog jezika, iz analize će se izuzeti pravila i problematika naglašavanja riječi[40].

Ad 1.) Pri izabiranju kriterija *jasnoća objašnjenja* polazilo se od pretpostavke da su objašnjenja u gramatikama za učenike koji nisu izvorni govornici hrvatskoga jezika jednoznačna, nedvosmislena te da pružaju jasne i točne odgovore na postavljena pitanja.

Ad 2.) Drugi je kriterij podrazumijevao *uporabljivost objašnjenja*, odnosno mogućnost neposredne primjene objašnjenja na rješavanje zadataka u udžbenicima hrvatskoga kao drugoga ili stranoga jezika bez korištenja sekundarnih izvora.

Imajući u vidu postavljene ciljeve analize jezičnih materijala, rezultati analize prikazat će se sa stajališta učenika hrvatskoga kao drugoga ili stranoga jezika s naglaskom na opis paradigmi onih vrsta riječi koje su, kako je u prethodnome poglavlju dokazano, najviše zastupljene u udžbenicima za učenje i poučavanje hrvatskoga kao stranoga jezika na njemačkome govornom području.

2.3.1. *Glagoli*

Sa stajališta učenja hrvatskoga kao drugoga ili stranoga jezika velikim se problemom postavilo usvajanje glagolske paradigme koja se temelji na njihovoj podjeli na glagolske vrste i razrede. Problem se odmah pojavljuje s nepreciznošću 7. glagolske vrste, koja sadrži nepravilne glagole. Naime, u PHG stoji: „Među nepravilne glagole mogli bismo ubrojiti mnoge glagole koje smo do sada spomenuli. Mnogi su od njih u različitim pojedinostima nepravilni, ali smo ih ipak uvrstili u neku od vrsta“ (Raguž 1997: 177).
Ako se pogleda glagolska paradigma prema vrstama u PHG (Raguž 1997: 146) i GHJ (Prikaz 4 i Prilog 2), vidi se da su se autori trudili paradigme što detaljnije predočiti, polazeći od pojašnjavanja tvorbi osnova.

[40] Važno je napomenuti da, iako se smatra da je „naglasak sastavni dio osnove, pa i on može sudjelovati u njezinoj promjeni. (…) Sve su takve promjene naglaska uvjetovane morfološkim zakonitostima, pa ih zato smatramo morfološkim, a ne fonološkim činjenicama“(GHJ 2005:42), u ovome se radu naglasak stavlja na računalnu morfološku analizu te se morfološki rječnik temelji na drugačijem načelu (Poglavlje 4).

174. **Peti razred**: *-r-Ø-ti /-ar-ē-m/-ar-Ø-ū: tr̂-Ø-ti / tȃr-ē-m/tȃr-Ø-ū.*

175. **Šesti razred**: *-lje-Ø-ti /-elj-ē-m/-elj-Ø-ū: mljȅ-Ø-ti / mȅlj-ē-m/mȅlj-Ø-ū.*

176. **Sedmi razred**: *-ije-Ø-ti / -Ø-ē-m/-Ø-Ø-ū: drijȅ-Ø-ti / drØ-ê-m/dr-Ø-û.*

177. **Osmi razred**: *-ije-//-es-Ø-ti /-es-ē-m/-es-Ø-ū: dȍ-nije-//-nes-Ø-ti / do-nès-ē-m / do-nès-Ø-ū.*

178. **Deveti razred**: *-je-Ø-ti / -i-(j)-ē-m / -ij-ū: smjȅ-Ø-ti / smȉ-(j)-ē-m / smȉ(j)-Ø-ū.*

179. **Deseti razred**: *-i-Ø-ti /-i(j)-ē-m / -i-(j)-Ø-ū: pȉ-Ø-ti / pȉ(j)-ē-m / pȉ(j)-Ø-ū -u-Ø-ti / -u(j)-ē-m / -u(j)-Ø-ū: čȕ-Ø-ti / čȕ(j)-ē-m / čȕ(j)-Ø-ū.*

180. **Jedanaesti razred**: *-g-Ø-ti / -g-Ø-/-ž-ē-: mȍg-Ø-ti / mȍg-Ø-u/mȍg-Ø-ū / mȏž-ē-š, -k-Ø-ti / -k-Ø-ū//-č-ē-m: rȅk-Ø-ti / rȅk-Ø-ū/rȅč-ē-m, -h-Ø-ti / -h-Ø ū /-š-ē-m: vȩ́h-Ø-ti / vȩ́h-Ø-ū/ vȩ́š-ē-m.*

181. **Dvanaesti razred**: *-ū-Ø-ti / -p-ē-m/-p-Ø-ū: pȍ-sū-Ø-ti / pȍ-sp-ē-m / pȍ-sp Ø-ū.*

182. **Trinaesti razred**: *-Ø-Ø-ti / -n-ē-m/-n-Ø-ū: stȃ-Ø-ti / stȃ-n-ē-m/stȃ-n-Ø-i -g-Ø-ti / -g-n-ē-m/-g-n-Ø-ū: mȏg-Ø-ti / mȏg-n-ē-m/mȏg-n-Ø-ū, -k-Ø-ti -k-n-ē-m / -k-n-Ø-ū: rȅk-Ø-ti / rȅk-n-ē-m/rȅk-n-Ø-ū.*

183. **Četrnaesti razred**: *-a-Ø-ti/-a-d-ē-m/a-d-Ø-ū: dȃ-Ø-ti / dá-d-ē-m/dá-d-Ø-; stȃ-Ø-ti / stȃ-d-ē-m/stȃ-d-Ø-ū.*

184. **Petnaesti razred**: *-a-Ø-ti /-a-dn-ē-m/-a-dn-Ø-ū: dȃ-Ø-ti / dȃ-dn-ē-m/dȃ-d Ø-ū, htjȅ-Ø-ti /htjȅ-dn-ē-m/htjȅ-dn-Ø-ū.*

185. **Šesnaesti razred**: *-d-Ø-ti /-d-ē-m/-d-Ø-ū: ȋd-Ø-ti / ȋd-ē-m//ȋd-Ø-ū.*

186. **Sedamnaesti razred**: *-īd-Ø-ti /-(ī)đ-ē-m/-(ī)đ-Ø-ū: nà-ī-Ø-ći / nà-īđ-ē-n nà-īđ-Ø-ū/nȃ-Ø-ći/nȃ-đ-ē-m/nȃ-đ-Ø-ū.*

187. **Osamnaesti razred:** *-e-Ø-ti /-Ø-ē-m/-Ø-Ø-ū: zrȅ-Ø-ti / zr-ê-m/zr-Ø-û.*

Prikaz 4. Izvadak iz prikaza glagolske paradigme u GHJ (Silić-Pranjković 2005: 46)

PHG navodi: „**Prezentska osnova**, naravno, vidi se iz prezentskih oblika bez nastavaka“ (Raguž 1997: 163). „**Infinitivna se osnova**, načelno**,** dobiva ako se odbaci infinitivni sufiks *ti*“ (Raguž 1997: 164). Nadalje, HG navodi: „Dosta glagola ima koji se tvore od infinitivne osnove po prvoj i drugoj vrsti, a prezent, imperativ i pridjev trpni po drugoj vrsti“ (Barić et al. 1995: 253). PHG specificira način tvorbe infinitive i prezentske osnove te među ostalim kaže: „Infinitivna osnova načelno je ono što se dobije ako se odbije infinitivni sufiks. To je tako u većini slučajeva, tj. kad je ispred infinitivnoga sufiksa *–ti* neki samoglasnik. (…) Ali u ostalim slučajevima, kad imamo suglasnik (*s*) ispred sufiksa *–ti* u infinitivu ili sufiks *–ći* (…) onda treba za svaku skupinu glagola znati koja je infinitivna osnova. Infinitivna je osnova potrebna u tvorbi *aorista, radnoga pridjeva, priloga prošloga* itd.“ (Raguž 1997: 163). U svojim primjerima PHG naznačuje posebno prezentsku osnovu masnim slovima:

Bosti – **bod**em
Krasti – **krad**em
Leći – ležem….**leg**u

GHJ kod svakoga glagola daje i infinitivnu i prezentsku osnovu (Prikaz 5) (Prilog 2), međutim ne navode se ostale osnove koje se upotrebljavaju u ostalim glagolskim vremenima.

Inf. osn.	Prez. osn.	1. l. jed.	2. l. jed.	3. l. jed.	1. l. mn.	2. l. mn.	3. l. mn.
pisati							
pīš-a-	*pīš-ē-//pīš-Ø-*	*pîš-ē-m*	*pîš-ē-š*	*pîš-ē-Ø*	*pîš-ē-mo*	*pîš-ē-te*	*pîš-Ø-ū*
srȅtati							
sret-a-	*sreć-ē-//sreć-Ø-*	*srȅć-ē-m*	*srȅć-ē-š*	*srȅć-ē-Ø*	*srȅć-ē-mo*	*srȅć-ē-te*	*srȅć-Ø-ū*
pozívati							
(po)zīv-a-	*(po)zīvlj-ē-// (po)zīvlj-Ø-*	*pòzīvlj-ē-m*	*pòzīvlj-ē-š*	*pòzīvlj-ē-Ø*	*pòzīvlj-ē-mo*	*pòzīvlj-ē-te*	*pòzīvlj-Ø-ū*
kázati							
kāz-a-	*kāž-ē-//kāž-Ø-*	*kâž-ē-m*	*kâž-ē-š*	*kâž-ē-Ø*	*kâž-ē-mo*	*kâž-ē-te*	*kâž-Ø-ū*
pljȕvati							
plju-va-	*plju-je-//plju-jØ-*	*pljȕ-jē-m*	*pljȕ-jē-š*	*pljȕ-jē-Ø*	*pljȕ-jē-mo*	*pljȕ-jē-te*	*pljȕ-Ø-jū*
brȉjati							
brȉ-ja-	*bri-jē-//brij-Ø-*	*brȉ-jē-m*	*brȉ-jē-š*	*brȉ-jē-Ø*	*bri-jē-mo*	*brȉ-jē-te*	*brȉj-Ø-ū*
pȅći							
pek-Ø-	*peč-ē-//pek-Ø-*	*pèč-ē-m*	*pèč-ē-š*	*peč-ē-Ø*	*pèč-ē-mo*	*pèč-ē-te*	*pèk-Ø-ū*
strȉći							
strig-Ø-	*strīž-ē-//strīg-Ø-*	*stríž-ē-m*	*stríž-ē-š*	*stríž-ē-Ø*	*stríž-ē-mo*	*stríž-ē-te*	*strig-Ø-ū*
vŕći							
vŕh-Ø-	*vŕš-ē-//vŕh-Ø-*	*vŕš-ē-m*	*vŕš-ē-š*	*vŕš-ē-Ø*	*vŕš-ē-mo*	*vŕš-ē-te*	*vŕh-Ø-ū*
žȅti							
že-Ø-	*žanj-ē//žanj-Ø-*	*žȁnj-ē-m*	*žȁnj-ē-š*	*žȁnj-ē-Ø*	*žȁnj-ē-mo*	*žȁnj-ē-te*	*žȁnj-Ø-ū*
klȇti							
klē-Ø-	*kun-ē-//kun-Ø-*	*kùn-ē-m*	*kùn-ē-š*	*kùn-ē-Ø*	*kùn-ē-mo*	*kùn-ē-te*	*kùn-Ø-ū*
klàti							
kla-Ø-	*kolj-ē-//kolj-Ø-*	*kȍlj-ē-m*	*kȍlj-ē-š*	*kȍlj-ē-Ø*	*kȍlj-ē-mo*	*kȍlj-ē-te*	*kȍlj-Ø-ū*

Prikaz 5: Izvadak iz opisa glagolske paradigme iz GHJ (Silić-Pranjković 2005: 62)

Za učenika hrvatskoga kao stranoga jezika ova se pojašnjenja i pregled paradigmi mogu učiniti nepreglednijima ili više zbunjujućima. Polazeći od navoda u PHG da „iako nema precizne općeprihvaćene klasifikacije glagola po vrstama, obično se dijele u šest vrsta, uz dodatak nepravilnih glagola“ (Raguž 1997: 63), razmišljalo se o podastiranju drugačijega rješenja i prikaza glagolske paradigme. Kao najprihvatljivijim se rješenjem nametnulo rješenje koje je korišteno pri izradi morfološkoga rječnika hrvatskoga jezika, a koje sadrži 96 uzoraka sprezanja koji su se koristili pri izradi HUMOR-a uzimajući u obzir samo računalno-morfološke karakteristike glagola u pisanim tekstovima (Poglavlje 7.4.). Jedan takav sličan postupak davanja uzoraka sprezanja napravio je Raguž u PHG, koji je prikazao uzorke sprezanja te u posebnom popisu abecednim redom naveo glagole koji su

„nositelji“ uzorka. Jedinom manom ovoga načina jest nedostatak popisa možda s najučestalijim glagolima koji bi se uklopili u neki od uzoraka.

Kao sljedeće pitanje koje se postavlja pri usvajanju glagolske paradigme jest pitanje glagolskoga vida, a time i glagolskih vremena imperfekta i aorista. Pri opisivanju pojave svršenosti ili nesvršenosti te njezine uporabe, GHJ govori: „Neki se glagolski oblici tvore od nesvršenih, a neki od svršenih glagola“ (Silić-Pranjković 2005: 58). PHG daje sljedeće objašnjenje: „Svaki glagol znači ili radnju, proces u razvoju (…) ili radnju kao završen proces. Prvi se zato zovu nesvršeni (imperfektivni glagoli, a drugi svršeni (perfektivni) glagoli“ (Raguž 1997: 157). Slično tomu, HG navodi sljedeće objašnjenje: „Glagoli koji izriču radnju u vršenju zovu se nesvršeni (imperfektivni) ili glagoli nesvršenoga (imperfektivnoga) vida. Glagoli koji izriču izvršenost radnje zovu se svršeni (perfektivni) ili glagoli svršenoga (perfektivnoga) vida“ (Barić et al. 1995: 225). Ako se ova objašnjenja analiziraju sa stajališta učenika, vidljivo je da se određivanje glagolskoga vida temelji i na semantici, što povlači za sobom cijelu glagolsku paradigmu. Možda bi se jednim od prihvatljivijih rješenja učinilo rješenje s čisto morfološkoga gledišta kao što je to navedeno u HUMOR-u (Poglavlje 7.4.2.). Slično rješenje nudi i GHJ koja glagolski vid i promjene opisuje s morfološkoga stajališta (Silić-Pranjković 2005: 52-53) te između ostaloga navodi: „Svršeni se glagoli tvore od nesvršenih bilo zamjenom sufiksalnoga morfema osnove nesvršenih glagola bilo pridruživanjem prefiksalnoga morfema osnovi nesvršenih glagola “(Silić-Pranjković 2005: 48). Polazeći s morfološkoga stajališta, i PHG daje slično objašnjenje, ali ne specificirajući one slučajeve kada se svršeni glagoli ne tvore od nesvršenih uporabom prefiksalnoga morfema: „Prefiksom se najčešće, ali ne uvijek, od nesvršenoga glagola može dobiti svršeni“ (Raguž 1997: 158).

Neprecizna objašnjenja predstavljaju dodatnu poteškoću pri usvajanju hrvatskoga kao drugoga ili stranoga jezika. Za učenike koji uče hrvatski jezik, primjerice, objašnjenje iz GHJ da „neki glagoli mogu biti i svršeni i nesvršeni. Takav je npr. glagol *vidjeti* (…), te većina glagola na *–irati*, npr. *telefonirati*, *lakirati,* ali *muzicirati, dirigirati* i sl. samo su nesvršeni“ (Silić-Pranjković 2005: 229), ne precizirajući iznimke i ne navodeći popis dvovidnih glagola, predstavlja dodatnu poteškoću za usvajanje hrvatskoga kao stranoga jezika. PHG slično tomu navodi: „Ipak ima **mnogo** glagola koji imaju oba vida (aspekta), to jest *dvovidni* su (*dvoaspektni)*. Mnogi od ovih dvovidnih glagola domaćeg, slavenskoga

podrijetla **ipak nisu u svim oblicima**, odnosno vremenima dvovidni, nego su u **nekim** vremenima jednoga vida, a u drugima drugoga vida“ (Raguž 1997: 158)[41].

Kao logičnim se slijedom postavlja pitanje o svrsi navedene problematike određivanja svršenih i nesvršenih oblika glagola. Naime, određivanje je glagolskoga vida temelj za glagolsku paradigmu opisanu u navedenim gramatikama, posebice za tvorbu aorista i imperfekta. Ako se podrobnije pogledaju opisi tvorbe ovih dvaju vremena, dolazi se do zaključka da u objašnjenjima manjka nužno specificiranje iznimaka, odnosno pozornost posvećena otklanjanju nedoumica za učenike hrvatskoga kao drugoga ili stranog jezika. Ovdje se može elaborirati i o pragmatičnoj strani aorista i imperfekta, odnosno o opravdanosti njihova uključenja u glagolsku paradigmu koju trebaju usvojiti učenici hrvatskoga kao drugoga ili stranoga jezika. PHG navodi da aorist „nije tako čest, ali je češći nego što se to tvrdi u hrvatskim gramatikama“ (Raguž 1997: 185). Nadalje, također napominje da je imperfekt „u suvremenome jeziku iščezlo glagolsko vrijeme. Postoji rijetko, samo u pojedinačnim slučajevima (od glagola *biti* najčešće)“ (Raguž 1997: 186) te da „pluskvamperfekt i aorist su sasvim rijetki, a imperfekt je praktično nestao iz upotrebe“ (160). Slično tvrdi i Tadić (1994) koji pri navođenju glagolskog oblika *zadronjaše* govori: „Svakome, tko iole bolje poznaje hrvatski, jasno je da oblik *zadronjaše* predstavlja ili treće lice množine aorista ili drugo i(li) treće lice jednine imperfekta nekoga glagola premda je teško pretpostaviti da bi tko mogao prepoznati o kojem se točno glagolu radi“ (Tadić 1994: 45). Opravdanost pak podučavanja aorista i imperfekta učenicima hrvatskoga kao drugoga ili stranoga jezika podupire i navod iz GHJ: „U suvremenoj se komunikaciji aorist zamjenjuje perfektom svršenih glagola. U novije se vrijeme, kao i imperfekt, počinje rabiti pri slanju poruke e-poštom, jer zauzima manje prostora od adekvatnoga glagolskog oblika za izražavanje prošlosti“ (Silić-Pranjković 2005: 96) te navod Ljudevita Jonkea (2005) koji govori: „Oduzmite našim liricima aorist, kao da ste im odrezali krila. Slično se može reći za naše epike s obzirom na imperfekt. To dakako nisu aoristi radi aorista, nego aoristi s izraženom funkcionalnošću i izmjenom s drugim vremenima, kad god je potrebno“ (Jonke 2005: 75).

Nedoumice koje se pojavljuju pri spomenu ovih glagolskih vremena odnose se u prvome redu na nedostatak relevantnih informacija u hrvatskim gramatikama sa stajališta učenja hrvatskoga kao stranoga jezika. Tako HG navodi da „aorist imaju svršeni i **rjeđe** nesvršeni

[41] Isticanja moja

glagoli“[42] (Barić et al. 1995: 238), dok za imperfekt stoji da ga imaju „samo nesvršeni glagoli“ (Barić et al. 1995: 239). Također se dalje u tekstu navodi da „**najviše** glagola tvori imperfekt alomorfom *a*. **Neki** glagoli mogu tvoriti imperfekt na dva načina…“[43] (Barić et al. 1995: 239). U navedenom se izvoru ne preciziraju *rijetki nesvršeni* glagoli koji mogu imati aorist, a isto se tako ne navode glagoli koji ne tvore imperfekt alomorfom *a*. Strancima koji počinju učiti hrvatski jezik ovakva se objašnjenja doimaju zbunjujućima, pogotovo što se u iznimke ubrajaju glagoli *piti, venuti, željeti, držati, brojiti, ljubiti, dati, imati, znati,* itd., koji se ubrajaju u skupinu najučestalijih glagola, a koji su ujedno i dvovidni (prema Barić et al 1995: 254ff). Jedini jezični priručnik koji je u svoju paradigmu uvrstio dvovidni glagol jest PHG, gdje taj glagol čini uzorak 51. paradigme. GHJ, međutim, ne izdvaja posebno one glagole koji imaju aorist, nego samo pojašnjava njegov način tvorbe, dok se u PHG pri objašnjavanju tvorbe aorista daje sljedeći navod: „Aorisni se nastavci dodaju obično svršenim glagolima, a rjeđe i nesvršenim glagolima. S nesvršenima ti nastavci daju značenje završenosti glagolima koji svojim vidom označuju trajanje radnje, dakle završenost radnje koja je trajala“ (Raguž 1997: 181).

Kao jedno od rješenja za zaobilaženje navedene problematike, može se ponuditi oslanjanje na primjerice RHJ, gdje je uz svaki glagol naveden i njegov vid, ili na GT. Na taj bi način učenik dobio nit vodilju za potrebitu glagolsku paradigmu i ne bi trebao sam pokušati odrediti glagolski vid. Problem se, međutim, pojavljuje kada se navodi u rječnicima ne podudaraju s paradigmama ili navodima u jezičnim priručnicima. Naime, HG kod prikaza glagolske paradigme glagola *tkati* daje i aoristne i imperfektne oblike, GT samo oblike za imperfekt, dok u RHJ stoji da je glagol *tkati* nesvršeni. Ovi navodi mogu stvoriti nedoumice pri usvajanju hrvatskoga kao stranoga jezika, jer se glagol *tkati* po učestalosti nalazi na 547. mjestu u hrvatskome jezičnom korpusu (prema HČR) od ukupno 569 mjesta te bi se time vjerojatno ubrojio u leksik koji se treba usvojiti na višim razinama s obzirom na to da mu je apsolutna čestoća 0.0023. Slično tomu, HG i HML navode aoristne oblike *stah* i *stadoh* od glagola *stati* (u HML ovi oblici imaju oznaku *Vmia1s*, koja pokazuje da se radi o istom obliku), dok GT i RHJ spominju samo jedan oblik, *stadoh*.

Ako se nadalje pomnije analiziraju primjeri kao uzorci paradigme u gramatikama i jezičnim priručnicima, postavlja se pitanje njihove reprezentativnosti. Primjerice, glagoli *strijeti*, *djeti,* koji se navode u HG, ne nalaze se u Anićevu rječniku. RHJ sadrži samo oblike glagola *djeti,* dok glagol *strijeti* nije uvršten u rječnik. Samim se time postavlja

[42] Isticanje moje
[43] Isticanja moja

pitanje autentičnosti Rječnika hrvatskoga jezika, odnosno broj lema koje se u Rječniku ne pojavljuju. Analizom sastavljenoga korpusa može se dobiti uvid u riječi koje nisu u sastavu rječnika, ali se pri tome postavlja pitanje jesu li one sastavnice hrvatskoga leksika.

Kod glagola *gnjiti* HG navodi dva oblika u prvom licu prezenta, *gnjim* i *gnjijem*, dok GT i HML navode samo jedan oblik: *gnjijem*. Još jedan od glagola koji se nalazi u HG, *gnati*, nije naveden u RHJ. Kao prvo lice prezenta jednine ovoga glagola HG navodi oblik *ženem*, dok je u GT isti glagol zastupljen, ali u obliku *gonim, goniš*..., a Anićev ga Rječnik ne sadržava. HML ga spominje u oblicima *gnam, gnaš gna*... itd. Zanimljivo je da se ovaj glagol nalazi na 555. mjestu od ukupno 569 mjesta u čestotnome rječniku te se u korpusu pojavljuje samo u oblicima *gnjije* (5 puta)*, gnjijem* (1 put)*, gnjiju* (4 puta)*, gnjila* (3 puta) i *gnjiti* (2 puta). Ako se nadalje, kao pokus, usporede primjerice glagoli *sresti* i *sjesti* koji su navedeni u PHG, opaža se sljedeće: PHG uz glagol *sresti* daje naznaku o njegovoj svršenosti. Te naznake nema u GT, kao ni oblika glagolskoga priloga *sretavši,* koji se nalazi u PHG. Umjesto toga oblika glagolskoga priloga, GT daje nam oblik *srevši.* Nadalje, PHG navodi dva oblika glagola *sresti* u prezentu jednine: *sretnem, sretneš*... i *sretem, sreteš*..., što primjerice nije navedeno u HML. Analogno tomu, PHG naznačuje i dva aoristna oblika: *sretoh*… i *sreh*, dok GT u oba slučaja ima samo jedan oblik. Podrobnije proučavajući paradigmu ovih glagola, uočava se da GT u imperativu navodi i imperativni oblik za prvo lice jednine *da sretnem*, što se ne nalazi u PHG. Kod glagola *sjesti* primjećuje se da je u GT za prezent naveden samo jedan oblik: *sjednem, sjedneš*..., dok PHG navodi još i oblik *sjedem, sjedeš*... Ako se dalje promotri paradigma ovoga glagola, primjećuje se da GT ima jedan aoristni oblik *sjedoh*, dok PHG navodi osim njega još i oblik *sjeh*. Slično ovom slučaju, GT spominje jedan oblik glagolskoga priloga prošlog: *sjevši*, dok se PHG koristi još i oblikom *sjednuvši*.

Osim razlika među navodima u različitim jezičnim priručnicima i gramatikama, postoje razlike i u navodima koji se odnose u prvome redu na pragmatičku uporabu riječi, ali koji se neizravno tiču i usvajanja morfoloških odrednica hrvatskoga standardnog jezika. Primjerice, u GT ne postoji navod da je oblik *velju* zastario oblik glagola *velim, veliš*, dok taj navod sadrži PHG: „Zastario i regionalan oblik postoji i od glagola *velim, veliš*…. (koji nema infinitiva) u obliku *velju*“ (Raguž 1997: 177).

Pri analizi opisa glagolske paradigme pojavilo se još jedno pitanje – tretiranje glagolske imenice i glagolskoga pridjeva, o čemu će više riječi biti u poglavlju 7.4. Dvojba se odnosi

na uključivanje navedenih kategorija ili u sastav glagolske paradigme, ili njihovo tretiranje kao kategorije imenica, odnosno pridjeva, navodeći njihovo podrijetlo. Raguž (1997) u svojoj gramatici ne spominje posebno kategoriju glagolske imenice, dok pri spomenu glagolskoga pridjeva ističe njegovo korištenje za tvorbu perfekta, pluskvamperfekta, obaju kondicionala i futura II., odnosno njegovu atributivnu, predikativnu i optativnu funkciju (Raguž 1997: 197). Pri izradi GENOBLIK-a, Tadić (1994) se koristio tradicionalnim rješenjem i navedene oblike uvrstio u sustav glagolske paradigme. Iz razloga što Anić u svome rječniku glagolsku imenicu i glagolski pridjev tretira kao zasebnu lemu samo u slučaju da ona ima i neko drugo značenje, nameće se zaključak da se i pri izradi morfološkoga rječnika glagolska imenica i glagolski pridjev tretiraju kao zasebne jedinice s imenskom i pridjevskom paradigmom navodeći njihovu polaznu lemu (glagol).

Uzimajući u obzir sve navedeno, dolazi se do zaključka da je za učenike koji uče hrvatski kao drugi ili strani jezik potrebno osmisliti drugačiji pristup podučavanju glagolske paradigme, a koji će se u prvome redu temeljiti na korpusnoj analizi, te koji će biti podrobnije predstavljen u Poglavlju 7.4.2.

2.3.2. *Pridjevi*

Ako se nadalje pomnije analizira pridjevska deklinacija, mogu se također uočiti razlike među pojedinim jezičnim priručnicima. U ovome će se poglavlju podrobnije prikazati rezultati analize gramatika i jezičnih priručnika sa stajališta deklinacije pridjeva, odnosno predznanja koja su potrebna učenicima kojima hrvatski nije materinski jezik da bi bili u mogućnosti uspješno deklinirati pridjeve u hrvatskome standardnom jeziku. Naime, za potrebe opisa pridjevske paradigme, odnosno deklinacije pridjeva, vidljivo je da pojedini autori podrazumijevaju semantičko predznanje. Već kod određivanja pridjevske paradigme potrebno je naime odrediti određenost, odnosno neodređenost pridjeva, te kategoriju živosti. Međutim, analizom jezičnih priručnika ispostavilo se da mnogi autori prvo polaze od pojašnjavanja podjele pridjeva i njihovih semantičkih karakteristika, da bi kasnije pojasnili razlikovanje pridjevskoga vida te se naposljetku bavili pridjevskom paradigmom. Taj će se model slijediti i ovdje, kod predstavljanja rezultata analize. Analiza se usredotočuje, kako je već bilo napomenuto, na morfološke karakteristike.

Za učenika hrvatskoga kao drugoga ili stranoga jezika prilično se zbunjujućim čini neujednačenost podjele pridjeva na kategorije, jer se ona neizravno veže uz pridjevski vid.

Primjerice, PHG pridjeve dijeli na *opisne pridjeve, odnosne pridjeve* i *posvojne pridjeve* (Raguž 1997: 88) te napominje da se kategorija određenosti i neodređenosti može pronaći samo kod opisnih pridjeva. GHJ, međutim, pri podjeli pridjeva spominje *kakvoćne (kvalitativne) pridjeve, posvojne (posesivne) pridjeve, gradivne (materijalne) pridjeve, odnosne pridjeve* (Silić-Pranjković 2005: 133f), koji se ne spominju u PHG. HG, nadalje, kod podjela pridjeva govori: „Pridjevi se po značenju mogu podijeliti na samo dvije skupine, i to na opisne i odnosne. Odnosni pridjevi (…) obuhvaćaju posvojne i gradivne“ (Barić et al. 1995: 174). „Određeni i neodređeni vid imaju samo opisni pridjevi“ (Barić et al. 1995: 179).

Nadalje, ako podrobnije analiziramo pojašnjenja u pojedinim gramatikama i jezičnim priručnicima koja se odnose na određivanje određenosti pridjeva, možemo pronaći sljedeće navode:

> Pridjevima se iskazuje i kategorija neodređenosti/određenosti. Tako se pridjevom *visok* u *visok stol* iskazuje neodređenost *stola*, a pridjevom *visoki* u *visoki stol* određenost *stola*. Po tome se pridjevi dijele na **neodređene pridjeve** i **određene pridjeve**. Neodređeni oblik pridjeva stoji uz neodređeni predmet iskazan imenicom, tj. uz predmet koji je sugovornicima u komunikaciji nepoznat, a određeni oblik – uz određeni predmet, tj. uz predmet koji je sugovornicima u komunikaciji poznat (Silić-Pranjković 2005: 134).
>
> Neodređeni pridjevi označuju kvalitetu, osobinu i odgovaraju na pitanje *kakav?*, a određeni pridjev označuje izabrani, utvrđeni (identificirani) primjerak, slučaj i odgovara pitanju *koji?* (Raguž 1997: 88).
>
> Po načinu kako određuju imenicu oblici pridjeva mogu biti neodređenog i određenog vida. Neodređeni vid kazuje promjenjive osobine onoga što znače imenice i odgovara na pitanje *kakav*. (…) Određeni vid upotrebljava se kad se pridjevom izriče stalna osobina predmeta, odnosno kad se određuje između više stvari različitih ona o kojoj se govori. Taj vid odgovara na pitanje *koji* (Barić et al. 1995: 174).

Ako se navedeni primjeri podrobnije analiziraju, zaključuje se da ovakav pristup zahtijeva već određeno poznavanje jezika, odnosno, ono što je sasvim razumljivo s obzirom na prirodu analiziranih jezičnih priručnika, učenike kojima je hrvatski materinski jezik.

Osim potrebne kategorije određenosti i neodređenosti, da bi se pridjevska paradigma u potpunosti mogla opisati, učenicima hrvatskoga kao stranoga jezika potrebna je još i kategorija živosti, što ponovno podrazumijeva upotrebu semantike i određivanje imenice

koja je pobliže određena pridjevom. Problem takvoga pristupa jest što se on ne zadržava više samo na području morfologije, nego zadire u područje semantike, odnosno kongruencije. Za učenike koji uče hrvatski kao strani jezik takav pristup podrazumijeva više razina. Prva je semantička razina, gdje učenik mora odrediti je li pridjev određeni ili neodređeni, zatim označuje li poslijepozicionirana imenica živo ili neživo, a druga razina podrazumijeva potom samu deklinaciju, promatranu s morfološkoga gledišta. Pri opisivanju pridjevske paradigme, do sada samo GHJ sklonidbu pridjeva promatra s potpuno morfološkoga stajališta, ne uzimajući navedene karakteristike u obzir te razlikuje deklinaciju pridjeva s i bez nepostojanog *a* (Prikaz 6) te navodi da „osnove nekih pridjeva u neodređenome obliku imaju nepostojano *a* u nominativu i akuzativu“ (Silić-Pranjković 2005: 134).

Pridjevi bez *nepostojanoga a* – određeni oblik

a) Muški rod – jednina	Muški rod – množina
N: *vèlik-ī*	*vèlik-ī*
G: *vèlik-ōga / vèlik-ōg*	*vèlik-īh*
D: *vèlik-ōmu / vèlik-ōme / vèlik-ōm*	*vèlik-īma / vèlik-īm*
A: (za živo) *vèlik-ōga / vèlik-ōg*	
(za neživo) *vèlik-ī*	*vèlik-ē*
V: *vèlik-ī*	*vèlik-ī*
L: *vèlik-ōmu / vèlik-ōme / vèlik-ōm*	*vèlik-īma / vèlik-īm*
I: *vèlik-īm*	*vèlik-īma / vèlik-īm*
b) Ženski rod – jednina	Ženski rod – množina
N: *vèlik-ā*	*vèlik-ē*
G: *vèlik-ē*	*vèlik-īh*
D: *vèlik-ōj*	*vèlik-īma / vèlik-īm*
A: *vèlik-ū*	*vèlik-ē*
V: *vèlik-ā*	*vèlik-ē*
L: *vèlik-ōj*	*vèlik-īma / vèlik-īm*
I: *vèlik-ōm*	*vèlik-īma / vèlik-īm*
c) Srednji rod – jednina	Srednji rod – množina
N: *vèlik-ō*	*vèlik-ā*
G: *vèlik-ōga / vèlik-ōg*	*vèlik-īh*
D: *vèlik-ōmu / vèlik-ōme / vèlik-ōm*	*vèlik-īma / vèlik-īm*
A: *vèlik-ō*	*vèlik-ā*
V: *vèlik-ō*	*vèlik-ā*
L: *vèlik-ōmu / vèlik-ōme / vèlik-ōm*	*vèlik-īma / vèlik-īm*
I: *vèlik-īm*	*vèlik-īma / vèlik-īm*

Prikaz 6: Uzorak pridjevske paradigme iz GHJ (Silić-Pranjković 2005: 136)

HG u svome tekstu pri sklonidbi pridjeva govori o „kvantiteti nastavaka“ (Barić et al. 1995: 175) jer uzima u obzir i naglaske.

JEDNINA	Rod		
	M	S	Ž
N	*-ø*	-o/-e	*-a*
G	*-a*		-ē
D	*-u*		-ōj
A	=G ili N	=N	-u
V	=N	=N	=N
L	*-u*		*-ōj*
I	*-īm*		-ōm

MNOŽINA		Rod	
	M	S	Ž
N	*-i*	-a	*-e*
G		*-īh*	
D		*-īm(a)/-ima*	
A	*-e*	=N	= N
V	= N	= N	= N
L		*-īm(a)/-ima*	
I		*-īm{a)/-ima*	

JEDNINA	Rod		
	M	S	Ž
N	*-ī*	-ō/-ē	*-a*
G	*-ōg(a)/- ēg(a)*		*-ē*
D	*-ōm(u/e)/- ēm(u)*		*-ōj*
A	=G ili N	=N	-ū
V	=N	=N	=N
L	*-ōm(e/u)/- ēm(u)*		=D
I	*-īm*		-om

MNOŽINA	Rod		
	M	S	Ž
N	*-i*	-ā	*-ē*
G	*-ih*		
D	*-īm{a/-ima*		
A	*-e*	=N	=N
V	=N	=N	=N
L	=D		
I	=D		

Prikaz 7: Uzorci pridjevske paradigme iz HG

Osim navedenih pojedinačnih tablica, HG daje i skupnu tablicu pridjevske paradigme (Prikaz 8):

		N			G			DL			A		
		m	s	ž	m	s	ž	m	s	ž	m	s	ž
jed-nina	neodr.	*-ø*	*-o*	*-a*	*-a*		-	*-u*		-	*-ø/-a*	-o	*-u*
	odr.	*-ī*	*-ō*	*-ā*	*-ōg{a)*		-	*-ōm(u/e)*		-	*-i/-ōg(a)*	*-ō*	*-ū*
mno žina	neodr.	*-i*	*-a*	*-e*	-			-			*-e*	*-a*	*-e*
	odr.	*-ī*	*-ā*	*-e*							*-ē*	*-ā*	*-ē*

Prikaz 8: Pridjevska paradigma iz HG

PGH (1997: 89) spominje dva tipa deklinacije, ovisno o različitim kategorijama: imeničku i zamjeničku deklinaciju, ali ih prikazuje u jednoj tablici (Prikaz 9).

		Jednina			Množina			Dvojina	
		M.	Sr.	Ž.	M.	Sr.	Ž.	M./Sr.	Ž.
Neodređeni	N	-Ø	-o/-e	-a	-i	-a	-e	-a	-e
	G	-a	-a	-ē	-īh	-īh	-īh		
	D	-u	-u	-ōj	-īm(a)	-īm(a)	-īm(a)		
	A	=N ili G	-o/-e	-u	-e	-a	-e		
	V	-Ø	-o/-e	-a	-i	-a	-e	-a	-e
	L	-u	-u	-ōj	-īm(a)	-īm(a)	-īm(a)	-a	-e
	I	-īm	-īm	-ōm	-īm(a)	-īm(a)	-īm(a)		
Određeni	N	-ī/Ø	-ō/-ē	-ā	-ī	-ā	-ē	-ā	-ē
	G	-ōg(a)/- ēg(a)		-ē	-īh	-īh	-īh		
	D	-ōm(u,e)/- ēm(u)		-ōj	-īm(a)	-īm(a)	-īm(a)		
	A	= N ili G	-ō/-ē	-u	-e	-ā	-ē		
	V	-ī	-ō/-ē	-ā	-ī	-ā	-ē	-ā	-ē
	L	-ōm(e, u)/- ēm(u)		-ōj	-īm(a)	-īm(a)	-īm(a)	-ā	-ē
	I	-īm	-īm	-ōm	-īm(a)	-īm(a)	-īm(a)		

Prikaz 9: Tablica koja prikazuje pridjevsku paradigmu iz PHG

Raguž (1997: 89) nadalje navodi da opisni pridjevi mogu imati i imeničku, i zamjeničku deklinaciju, ali da „imenička deklinacija, međutim, nije potpuno imenička; u njoj je izmiješan sustav imeničke i zamjeničke deklinacije“ (Raguž 1997: 89). Sa stajališta učenika ovakva se objašnjenja doimaju prilično zbunjujućima. Jedno od rješenja koja su se

nametnula pri izradi morfološkoga rječnika odnosno HUMOR-a prikazana su u poglavlju 7.3., a sadržavaju računalno-morfološke kategorije stemova i termova.

Podrobnije promatrajući navode u gramatikama koji opisuju samu pridjevsku paradigmu, odnosno konkretne oblike u pojedinim padežima, mogu se pronaći slični navodi koji učenicima predstavljaju teže usvajanje pridjevske paradigme, odnosno analizu već postojećih oblika. I PHG djelomice priznaje poteškoće pri usvajanju deklinacije pridjeva, te navodi: „Zbog nekih istih padeža i za određene i za neodređene pridjeve (npr. *dobrim ljudima / dobrim čovjekom, dobrom ženom* i sl.) iz pisanoga se teksta, pa ni iz izgovorenoga, ne može utvrditi o kojem je, kakvome pridjevu riječ“ (Raguž 1997: 89). Uspoređujući nadalje leme u RHJ, vidljivo je da navodi u rječnicima ne pridonose lakšem usvajanju sklonidbe pridjeva, odnosno ne olakšavaju usvajanje materijala opisanih u gramatikama i jezičnim priručnicima. GHJ primjerice navodi da „ima međutim pridjeva koji nemaju određenoga oblika i pridjeva koji nemaju neodređenog oblika“ (Silić-Pranjković 2005: 134). U RHJ primjerice kod lema *mačji* i *hrvatski* nije navedena činjenica da se ovdje radi o određenom obliku pridjeva koji nema svoga neodređenoga parnjaka. Iako se možda ovaj navod čini trivijalnim, sa stajališta učenika kojima hrvatski nije materinski jezik, ovaj podatak dobiva na važnosti.
Ako se nadalje pogleda stupnjevanje pridjeva, uočava se da GHJ teži k točnosti navoda te izbjegava objašnjenja poput: „neki dvosložni pridjevi (…) neki jednosložni pridjevi“ (Barić et al. 1995: 182). Primjerice, GHJ navodi sve one pridjeve koji ne tvore komparativ morfemom *-j-* (Silić-Pranjković 2005: 139). Promatrajući navode koji se odnose na stupnjevanje pridjeva, može se uočiti da primjerice Anić u svome rječniku u pravilu kod nepravilnih oblika navodi oblike za komparativ i superlativ (primjerice kod leme *dobar*), dok kod pravilnih oblika ne daje nikakve informacije o komparativu i superlativu. GT navodi komparativ kod oblika *crven* (oznaka a1), kod oblika *crveni* (oznaka a11) daje navod da on nema komparativ, dok RHJ kod leme *crven* ne navodi oblike za komparativ i superlativ. HML određeni oblik ne svrstava u zasebnu lemu, nego oblik navodi kao dio paradigme neodređenoga oblika *crven*.

Kao što se kod glagolske paradigme pojavila problematika tretiranja glagolske imenice i glagolskoga pridjeva u RHJ, odnosno time i u leksičkoj bazi HUMOR-a, i ovdje se postavlja pitanje vjerodostojnosti navoda u Anićevu rječniku sa stajališta statističke obrade. Naime, u rječnik su uvršteni samo neki posvojni pridjevi (primjerice *bratov*, ali ne

i *sestrin*, *mamin, tatin*, *stričev*) te se time postavlja pitanje može li se uopće govoriti o broju pridjeva koji su sastavni dio hrvatskoga leksikona te o broju lema koje RHJ sadržava. Analogno tomu postavlja se pitanje određivanja onih lema koje bi se mogle ubrojiti u dio osnovnoga (temeljnoga) hrvatskoga rječnika, ili bi se moralo u tome slučaju u potpunosti osloniti na korpusnu analizu.

2.3.3. *Imenice*

Kada se podrobnije promotri kategorija imenica u analiziranim gramatikama, jezičnim priručnicima i Anićevu rječniku (2000), kao prvi problem pojavljuje se odsutnost ili dvojbenost informacija koje se mogu pronaći, a koje su osnovica za cjelokupnu sklonidbu. Prvi uočeni problem jest kategorija gramatičkoga roda, o čemu će više govora biti u poglavlju 8.4.2., posebice u slučajevima kod zbirnih imenica. Naime, dok je u RHJ imenica *vrata* označena kao srednji rod, pluralia tantum, naznačujući time određenu paradigmu, GT za istu imenicu daje navod ženski rod i paradigmu samo u množini. Iako se time da zaključiti da je imenica pribrojena u kategoriju pluralia tantuma, za učenike kojima hrvatski nije materinski jezik, različito navođenje roda kod pojedinih imenica s pravom ih može zbuniti. Isto tako, ako pogledamo druge riječi koje pripadaju ovoj skupini, primjerice *prsa, pleća* i *leđa*, vidimo da se radi o različitim rodovima. RHJ za *leđa* navodi kategoriju srednjega rod, pluralia tantum, za *prsa* ženski rod pluralia tantum, a za *pleća* srednji rod množine. Ako se pogleda paradigma ovih imenica, vidljivo je da one imaju iste oblike, što dovodi do zaključka da bi se ipak radilo o istome gramatičkome rodu. Ako se ove imenice provjere u GT, može se vidjeti da GT svih triju imenica stavlja kategoriju ženskoga roda, ponovno ne naznačuje da se radi o imenicama koje imaju samo množinu, nego je iz paradigme vidljivo da nedostaje jednina. Isto tako oznaka u GT jednaka je za sve tri riječi (*n691*), iz čega se da zaključiti da navodi u RHJ nisu opravdani.

Ako se nadalje pogledaju zbirne imenice, uočavaju se neke razlike između RHJ i GT. U RHJ, primjerice, za imenicu *momčad* stoji da je imenica zbirna, da dolazi od imenice *momče.* GT ju, međutim, svrstava u dvije kategorije. U jednoj se *momčad* pojavljuje kao imenica koja nema svoju množinu, dakle kao zbirna imenica ili singularia tantum, a u drugoj se verziji ona pojavljuje kao množina od imenice *momče.* Ovakve dvije vrste

tretiranja navedene imenice dovode do diskrepancije, a time i do nejasnoća pri učenju hrvatskoga kao drugoga ili stranoga jezika.

Ponovno se i kod imenica pojavljuje problem primjera koji nisu navedeni u RHJ te time i pitanje relevantnosti reprezentativnih paradigmi za učenje hrvatskoga kao stranoga jezika. Primjerice, GHJ navodi kao reprezentativan primjer paradigme imenicu *golocijevka*, koja nije sadržana u RHJ.

Ako se pogleda morfološki opis imenske paradigme, uočavaju se bitne razlike u različitim pristupima. GHJ opisivanju sklonidbe imenica pristupa samo s morfološkoga stajališta i uzima u obzir samo razlikovne morfeme te tipove deklinacija razlikuje samo prema alternacijama osnove (Silić-Pranjković 2005: 98ff), kao što se vidi iz prikaza 10.

IMENICE SREDNJEGA RODA				
Jednina				
N: *sèl-o*	*pȍlj-e*	*plȅm-e-Ø*	*dijét-e-Ø*	*sédl-o*
G: *sèl-a*	*pȍlj-a*	*plȅm-en-a*	*djèt-et-a*	*sédl-a*
D: *sèl-u*	*pȍlj-u*	*plȅm-en-u*	*djèt-et-u*	*sédl-u*
A: *sèl.o*	*pȍlj-e*	*plȅm-e-Ø*	*dijét-e-Ø*	*sédl-o*
V: *sèl-o*	*pȍlj-e*	*plȅm-e-Ø*	*dijét-e-Ø*	*sédl-o*
L: *sèl-u*	*pȍlj-u*	*plȅm-en-u*	*djèt-et-u*	*sédl-u*
I: *sèl-om*	*pȍlj-em*	*plȅm-en-om*	*djèt-et-om*	*sédl-om*
Množina				
N: *sȅl-a*	*pòlj-a*	*plem-èn-a*	*(djèc-a* - kao *žèn-a* u jednini)	*sédl-a*
G: *sȇl-ā*	*pólj-ā*	*plem-én-ā*		*sedál-ā/ sȅdāl-ā/ sédl-ā*
D: *sȅl-ima*	*pòlj-ima*	*plem-èn-ima*		*sédl-ima*
A: *sȅl-a*	*pòlj-a*	*plem-èn-a*		*sédl-a*
V: *sȅl-a*	*pòlj-a*	*plem-èn-a*		*sédl-a*
L: *sȅl-ima*	*pòlj-ima*	*plem-èn-ima*		*sédl-ima*
I: *sȅl-ima*	*pòlj-ima*	*plem-èn-ima*		*sédl-ima*

Prikaz 10. Uzorak paradigme deklinacije imenica srednjega roda iz GHJ

Jedina kategorija koju uzimaju u obzir jest kategorija žive ili nežive imenice, što za sobom povlači pitanje semantike. RHJ, naime, kod lema ne sadrži oznaku radi li se o živoj ili

neživoj imenici, tako da se ovakav pristup može učiniti problematičnim sa stajališta učenika kojima hrvatski nije materinski jezik. Oni se moraju, prije nego što započnu sklonidbu imenica, slijedeći Silić-Pranjkovićev pristup, prvo pozabaviti značenjem riječi, a tek onda morfološkim oblicima koje ta riječ dobije tijekom deklinacije.

HG pojašnjava da se „nastavci dodaju na osnovu, koja se, u pravilu, dobije ako se u gen. jedn. izostavi nastavak“ (Barić et al 1995: 104), što može dovesti do nejasnoća jer u RHJ ne postoje navodi za genitiv jednine svih imenica, te učenik kojemu hrvatski nije materinski jezik ne može ništa započeti s ovakvom vrstom definicije.

PHG i HG imaju drugo polazište. Težište opisa imenske paradigme u ovim udžbenicima leži naime u razlikama među trima tipovima deklinacije. Raguž ih naziva *a*-deklinacija, *e*-deklinacija i *i*-deklinacija, prema sufiksima koje imenice dobiju u genitivu jednine, a HG *vrsta a, vrsta e* i *vrsta i* (Prikazi 11, 12 i 13).

	Muški rod		Srednji rod	
	Jednina	Množina	Jednina	Množina
N	-Ø, -o/e	-i	-o/e	-a
G	-a	-ā, -ī, -ijū	-a	-ā
D	-u	-ima	-u	-ima
A	=N/=G	-e	-o/e	-a
V	-e/u, =N	-i	-o/e	-a
L	-u	-ima	-u	-ima
I	-om/em	-ima	-om/em	-ima

Prikaz 11: Uzorak *a-deklinacije* iz PHG

	Jednina	Množina
N	-a, -o, -e	-e
G	-ē	- ā, - ū, - ī
D	-i	-ama
A	-u	-e
V	-a, -e, -o	-e
L	-i	-ama
I	- ōm	-ama

Prikaz 12: Uzorak *e-deklinacije* iz PHG

	Muški rod		Srednji rod	
	Jednina	Množina	Jednina	Množina
N	-Ø, -o/e	-i	-o/e	-a
G	-a	-ā, -ī, -ijū	-a	-ā
D	-u	-ima	-u	-ima
A	=N/=G	-e	-o/e	-a
V	-e/u, =N	-i	-o/e	-a
L	-u	-ima	-u	-ima
I	-om/em	-ima	-om/em	-ima

Prikaz 13: Uzorak *i-deklinacije* iz PHG

Ako se podrobnije analizira njegov pristup sklonidbi imenica, uočava se da Raguž imenice dijeli prema njihovim gramatičkim rodovima i nastavku u genitivu jednine. Već se na prvi pogled može uočiti da u tablicama sa stajališta učenika kojima hrvatski nije materinski jezik nedostaje preglednosti, odnosno da učenik bez potrebitoga materinskog predznanja jezika teško može pravilno sklanjati riječ oslanjajući se samo na tablice iz PHG. Ova izjava proizlazi iz činjenice da je korištenje spomenutih tablica moguće samo onda kada se ne događaju promjene u osnovama. Iako autor tvrdi da

„Prema obliku nom. jedn. vrsta **a** muškoga roda dijeli se na dvije skupine:
1. na imenice koje u tom padežu imaju nastavak *–Ø*, i
2. na imenice koje u tom padežu imaju nastavak *–o, -e*“ (Raguž 1997: 104),

iz tablica se ne može lako donijeti zaključak koji bi išao u prilog ovoj tvrdnji.

Ako se nadalje tablica za a-deklinaciju iz PHG usporedi s tablicom a-deklinacije iz HG (Prikaz 14), uočavaju se neke razlike:

Padež	Jednina		Množina	
	m. r.	sr. r.	m. r.	sr.
N	*-Ø,-o/-e*	*-o/-e,-Ø*	*-ī*	*-a*
G	*-a*	*-a*	-ā, *-ī*	-ā
D	*-u*	*-u*	*-ima*	*-ima*
A	=N ili G	=N	*-e*	=N
V	*-e/-u* =N	=N	=N	=N
L	=D	=D	=D	=D
I	*-om/-em*	*-om/-em*	=D	=D

Prikaz 14: Tablica koja prikazuje a-deklinaciju iz HG

Naime, kao što se na prvi pogled može uočiti, za razliku od PHG, HG ne spominje nastavak *–iju*, koji je naveden u množini muškoga roda a-deklinacije u PHG. Osim toga, za razliku od HG, PHG ne navodi činjenicu da se neki oblici podudaraju s drugima (primjerice u akuzativu, vokativu i lokativu srednjega roda), nego upisuje iste oblike. S te je strane rješenje u HG praktičnije.

Osim toga, navodi iz HG ne pridonose lakom usvajanju hrvatskoga kao drugoga ili stranoga jezika, odnosno jednoznačnosti u pogledu sklonidbe. Tamo se naime navodi; „po vrsti *a* sklanjaju se imenice muškoga i srednjeg roda, po vrsti *e* **većinom** imenice ženskoga roda i **neke** muškoga, a po vrsti *i* imenice ženskoga roda[44]“ (Barić et al. 1995: 103). Osim toga, sve su imenice podijeljene na jednakosložne (jednak broj slogova u svim padežima) i nejednakosložne (nejednak broj slogova u svim padežima).

Kao najveći problem kod imenske paradigme pokazao se oblik u vokativu jednine, što je podrobnije pojašnjeno u poglavlju 8.4.1. U jezičnim priručnicima, naime, ne postoje konkretni navodi o tome koji se nastavak koristi isključivo kod kojih imenica te se najveći problem pojavljuje kada se otkriju nesuglasice između pojedinih priručnika (više o tome u poglavlju 8.4.2). Osim vokativa, problem nesuglasica između jezičnih priručnika prisutan je i kod drugih oblika. Naime, neki oblici koji se spominju u gramatikama nisu navedeni u RHJ. Tako primjerice RHJ navodi oblike *trijesaka / trijeski* za G mn. imenice trijeska, dok GHJ spominje samo oblik *trijesaka* s dvama naglasnim tipovima.

Što se tiče sufikasa koji se rabe u množini, HG navodi sljedeće pravilo: „jednosložne osnove obično se u množini produžuju množinskim umetkom *–ov-,* npr. *dom – dom-ov-i, drug- drug-ov-i, sin- sin-ov-i*. Te imenice mogu imati i kratku množinu (*domi, druzi, sini*), ali takav oblik većinom ima stilsku vrijednost pa se uglavnom upotrebljava u umjetničkoj prozi i u poeziji, a manje ili nikako u običnom govoru“ (Barić et al. 1995: 106).

Analizom imenske paradigme u jezičnim priručnicima koji se koriste u nastavi hrvatskoga kao drugog ili stranog jezika došlo se do zaključka da su učenicima kojima hrvatski nije materinski jezik potrebne konkretnije informacije i formalističko rješenje koje će im dati odgovore na sva postavljena pitanja. Jedno takvo rješenje moguće je obradom svih lema iz Anićeva rječnika pri izradi morfološkoga analizatora za hrvatski standardni jezik, o čemu će biti govora u sljedećim poglavljima ovoga rada.

[44] Isticanja moja.

2.3.4. *Zamjenice i brojevi*

Za primjer deklinacije zamjenica Raguž (1997) u svojoj gramatici daje svojstvenu tablicu za sklanjanje zamjenica, koja se ne nalazi u ostalim analiziranim gramatikama. Težeći jednostavnosti, on je pokušao ujediniti paradigmu i prikazati sustav koji se temelji na morfološkim odrednicama hrvatskoga standardnog jezika.

	M. rod	Sr. rod	Ž. rod
Jednina			
N	-Ø, - āj/- ī	-ē/-e, -ō/-o	-ā/-a
G	-ōg(a)/-og(a), -ēg(a)/-eg(a), -og(ā)		-ē
D	-ōm(u,e)/-om(u,e), -ēm(u)/-em(u)		-ōj
A	= N/G	= N	-ū/u
V	-Ø/-āj/-ī	= N	-ā /a
L	-ō m(e,u)/-om(e,u), -ē m/-em		-oj
I	-īm(e)		-om
Množina			
N	-ī/-i	-ā/-a	-ē/-e
G	-īh/-iju	-īh/iju	-īh/-iju
D	-īm/īma	-īm/īma	-īm/īma
A	-ē/-e	-ā/-a	-ē/-e
V	-ī/-i	-ā/-a	-ē/-e
L	= D	= D	= D
I	= D	= D	= D

Prikaz 15: Tablica padežnih nastavaka zamjeničkog tipa iz PHG

Za razliku od PHG, GHJ sklonidbu zamjenica predstavlja sa zasebnim paradigmama, dok HG daje jednu zajedničku tablicu (Prikaz 16).

	Jednina		
	m. r.	sr. r.	ž. r.
N	*- Ø, - āj, - ī*	*-o, - ō//-e, -ē*	*-a, - ā*
G	*-og(a)/- ōg(a), -eg(a)/- ēg(a)*		*-ē*
D	*-omu, -om, -ome, -emu(-ēmu), -em(-ēm)*		*-ōj*
A	=N ili G	=N	*-u, -ū*
V	=N		=D
L	=D		=D
I	*-īm, -ime*		*- ōm*
	Množina		
	m. r.	sr. r.	ž. r.
N	*-i*	*-a*	*-e*
G		*-īh, -ijū*	
D		*-ima, -īm*	
A	*-e*	=N	=N

V	=N
L	=D
I	=D

Prikaz 16: Uzorak zamjeničke paradigme iz HG (Barić et al 1995: 210)

Na pitanje koji se pristup pokazao učinkovitijim kod učenika kojima hrvatski nije materinski jezik, odgovor bi moglo dati samo jedno pilot istraživanje.

Nadalje, kod prikaza kategorije brojeva, PHG brojevne riječi dijeli na brojevne imenice, brojevne pridjeve, brojevne priloge i priložne izraze (Raguž 1997: 104) te posebno navodi svaki tip sklonidbe. Isto tako PHG podrobnije objašnjava i kritizira tvrdnje koje se nalaze u drugim gramatikama:

> Tako će biti pogrešno ono što se može naći u mnogim hrvatskim gramatikama i primjerima kad je riječ o upotrebi padežnih oblika za brojevne riječi tipa *petero, pet* itd., npr. kad treba upotrijebiti neki padež od sintagme *pet konja, sedam žena* i sl., pa se daju oblici *peterim konjima, sedmerim ženama* i sl. Te su sintagme načelno pravilne, ali ne za *pet konja i sedam žena*, nego za *peteri konji* i *sedmere žene* (što bi značilo „pet skupina konja" i „sedam skupina žena" neodređenoga broja). Za *pet konja* i *sedam žena* (i sl.) nema padežnoga oblika pa se za izražavanje padežnoga odnosa mora pribjeći drugačijoj konstrukciji rečenice, npr. *Bilo je pet konja, sedam žena* (i sl.) *kojima*… (Raguž 1997: 114)

Kao zaključak provedene analize, može se reći da je potrebno osmisliti novi način prikazivanja gramatičkih sadržaja učenicima kojima hrvatski nije materinski jezik, što će se djelomice i pokušati pri izradi morfološkoga rječnika koji će se temeljiti na morfološkoj analizi. Detalji o problemu sastavljanja morfološkoga rječnika za podučavanje hrvatskoga kao stranoga jezika bit će predočeni u poglavlju 3.

2.3.5. *Analiza Rječnika hrvatskoga jezika*

U nastavi hrvatskoga kao stranoga jezika prema *Kurikulumu* postavljeni su sljedeći ciljevi: „Učenici/učenice znaju i upotrebljavaju rječnike i priručnike hrvatskoga jezika" (Kurikulum 2008: 85), „samostalno i aktivno upotrebljavati jezične priručnike, rječnike i općeobrazovnu literaturu" (Kurikulum 2008: 45), „upoznati jezične priručnike i rječnike hrvatskoga jezika" (Kurikulum 2008: 52). Iz ovoga razloga smatram potrebnim u ovoj fazi rada predstaviti Anićev (2000) *Rječnik hrvatskoga jezika* na CD-romu sa stajališta učenika

hrvatskoga kao drugoga i stranoga jezika, usredotočujući se, sukladno ciljevima ovoga rada, na morfološke karakteristike riječi. Anićev je rječnik izabran još iz razloga što on čini osnovicu morfološkoga analizatora te morfološkoga rječnika, kao što je podrobnije predstavljeno u poglavlju 4.2.

Za analizu će se izabrati kriteriji koje je primijenio Prószéky (1997b) u svome radu. On naime ocjenjuje rječnike prema dvadesest parametara. Od njegovih će se tema izabrati sljedeći kriteriji za ocjenjivanje RHJ:

1. Učenik ne zna točan oblik riječi, treba mu rječnik.
2. Učenik ne zna koja je lema zadane riječi.
3. Učenika zanima značenje riječi.
4. Učenik želi znati nešto o etimologiji riječi.
5. Učenika zanima kako se riječ rastavlja.

Ad1.

Ako učenika zanima točan oblik riječi, pri korištenju RHJ nudi mu se izbor samo statičnoga pretraživanja rječnika, iako se možda on njime koristi na CD romu. Za razliku od ostalih suvremenijih dinamičnih rječnika, RHJ na CD romu služi poput papirnatoga izdanja te nudi pretraživanje samo prema lemama. Suprotno tomu, GT nudi opciju upisivanja oblika riječi te mu pretraživač kasnije ponudi odgovarajuće leme. Ako učenik pronađe odgovarajuću lemu, rječnik mu ne nudi sve oblike, nego samo neke, što znači da je osim rječnika učeniku potrebna i gramatika. U gramatikama, kao što se moglo pročitati, ne postoje opisane paradigme svih riječi, što nas dovodi do zaključka da je učeniku potreban morfološki rječnik sa svim oblicima i mogućnošću dinamičnoga pretraživanja rječnika prema oblicima, lemama, a po potrebi i stranome jeziku (više o tome u poglavlju 8.1.).

Ad2. Nakon što učenik na temelju drugih izvora dođe do saznanja o tome koja lema pripada traženom obliku, upisuje lemu u tražilicu rječnika. Kursor mu se lijeve strane zaustavlja na traženoj riječi na dnu ekrana, što je nezgodno, jer se riječ ne vidi u potpunosti te učenik mora pomicati prozor da bi došao do željene leme. Hrvatska rječnička tradicija u hrvatskome jeziku nalaže određene oblike riječi kao rječničke oblike. Sa stajališta učenja hrvatskoga jezika možda bi se praktičnijim pokazalo korištenje nekih drugih oblika kod pojedinih vrsta riječi (primjerice kod pridjeva), što će biti pojašnjeno u poglavlju 7.3.

Ad3. Rječnik nudi iscrpne informacije o značenjima riječi. Zanimljivo je da obiluje natuknicama s područja botanike (imena rijetkih vrsta biljaka i njihovih regionalnih inačica), a koje nisu od prevelike važnosti za prosječnoga govornika jezika.

Ad4. Rječnik nudi informacije o podrijetlu pojedinih riječi u obliku navoda jezika iz koje riječ potječe (primjer riječi stranoga podrijetla, poput *hiperonim*, *Hindu*, *brio*). Dinamičnije mogućnosti njegova pretraživanja poput mogućnosti koje nudi MobiDic (poglavlje 9.1.1.) ovdje nisu moguće.

Ad5. Učenik u ovome rječniku neće dobiti informacije o rastavljanju riječi na slogove. Takva vrsta informacije dostupna je primjerice govornicima njemačkoga jezika, gdje u Dudenovu (1997) rječniku okomite crtice označavaju mogućnost rastavljanja riječi. Informacije o rastavljanju riječi dostupne su preko programa za rastavljanje riječi, tzv. *hyphenatora*.

Na temelju ove analize došlo se do zaključka da učenicima koji uče hrvatski kao drugi ili strani jezik nije dovoljan samo RHJ kao referentna literatura, nego bi bilo poželjno imati i neku drugu vrstu rječnika, u najboljem slučaju dinamični rječnik s više funkcija kako je opisano u Prószékyjevu (1997b) radu. Dok se takav rječnik ne usavrši, veliku će pomoć pri učenju hrvatskoga jezika učenicima ponuditi računalni morfološki rječnik čiji će način rada i ustrojstvo biti predstavljeni u sljedećim poglavljima.

3. Morfološka analiza u službi izrade morfološkoga rječnika

3.1. Načelo izrade morfološkoga rječnika hrvatskoga standardnog jezika

Prije same izrade morfološkoga rječnika hrvatskoga standardnog jezika pojavilo se nekoliko nedoumica i dvojbi koje su se morale definirati. Kao što je već prije bilo napomenuto, morfološki je rječnik hrvatskoga jezika nastao kao jedan od proizvoda rada na morfološkoj analizi hrvatskoga standardnoga jezika, odnosno smatra se jednim od dvaju računalnih modela koji služe podučavanju hrvatskoga kao stranog jezika, a koji se temelje na morfološkoj analizi. Prije opisa same izrade morfološkoga rječnika te nadalje davanja teorijskoga okvira za izradu hrvatsko-njemačko-hrvatskih prevoditeljskih softvera, potrebno je objasniti razloge odabiranja procesa morfološke analize, a ne sinteze kao polazišnoga procesa.

Već je i Marko Tadić (1994: 31) u svojoj doktorskoj disertaciji posvetio poglavlje računalnoj morfološkoj sintezi, odnosno analizi. Između ostalog navodi:

> Premda postoje računalni sustavi koji se bave obradom morfologije prirodnih jezika u oba smjera - sintezom i analizom, barem su dva razloga zašto je GENOBLIK[45] ograničen samo na sintezu. Prvi je kompleksnost morfološke analize tj. „prepoznavanja“ pripadnosti pojedinih oblika riječi iz korpusa njihovim paradigmama. Ta analiza nužno mora uključivati elemente sintakse u slučajevima npr. složenih glagolskih vremena, neodređenih zamjenica razdvojenih enklitikama ili složenih brojeva tim prije što ti oblici često ne stoje u izravnom dodiru već su više ili manje udaljeni unutar iste ravnine. Drugi je razlog taj što je GENOBLIK prvenstveno zamišljen za provjeru istraživačkih hipoteza, a njezin je smjer, barem za sada, odozgo prema dolje što je smjer sinteze tj. proizvođenja, a ne prepoznavanja oblika (Tadić 1994: 31-32)

Kao što se može pročitati, 1994. je godine Tadić smatrao da je razvijanje programa koji bi se temeljio na morfološkoj analizi izrazito složeni proces, jer se pojavljuju tehnička ograničenja koja se tiču granica analize, odnosno korištenja primjerice znaka *space* kao graničnika. Taj je problem danas djelomice riješen jer je primjerice morfološki analizator HUMOR obogaćen i sintaktičkim kategorijama te se u takvom obliku koristi pod nazivom HumorEsk (više o tome u poglavlju 4.). HumorEsk je u mogućnosti danas prepoznavati i analizirati složena vremena te je jasnim definiranjem sastavnica riješeno pitanje

[45] GENOBLIK je ime računalnoga programa koji je osmislio Tadić, a pod tim imenom podrazumijeva i jezikoslovni model, i njegovu izvedbu (Tadić 1994: 7).

ograničavanja područja analize. Međutim, iz razloga što je svrha ovoga rada opis procesa izrade morfološkoga rječnika, nije se odlučilo za verziju HumorEsk, nego se na temelju HUMOR-a izradio kako leksički, tako i gramatički temelj rječnika i programa, te se cijeli proces morfološke analize zadržao na razini lema. Time se poimanje procesa morfološke analize razlikuje od opisanoga poimanja, u kojemu Tadić pod tim nazivom podrazumijeva vjerojatno analizu koja se temelji na tradicionalnim kategorijama, a koja uključuje i elemente sintakse.

Ako iscrpnije promotrimo Hrvatski morfološki leksikon[46] (Tadić 2003: 41), možemo uočiti da se on temelji na načelu morfološke sinteze, odnosno generiranja oblika, te da je nastao unutar projekta koji je započeo Marko Tadić 1994. godine. Pri obrazloženju izbora procesa generiranja autori polaze od generativne gramatike te govore:

> Ideja generiranja svih mogućih kombinacija morfema prvi se puta pojavila kod Hallea (1973) unutar generativne gramatike. Iako se tada činilo da se morfologija može prikazati kao jedinstvena podsastavnica leksikona u generativnoj gramatici, ta je koncepcija povukla nekoliko teoretskih problema. (…) Naša zamisao modeliranja morfologije hrvatskoga bila je na tome tragu, ali smo željeli model ostaviti koliko god je moguće jednostavnim i koristiti se računalnim podatcima iz istraživanja hrvatskoga jezika koje je do sada bilo provedeno[47] (Tadić-Fulgosi 2003: 41).

Nadalje, opisujući dijelove procesa sinteze, odnosno generiranja oblika, autori se pozivaju na Tadićev projekt iz 1994. godine i navode:

> Fleksija: modelirana kao popis stemova koji se generiraju i popis fleksijskih morfema s pravilima za njihovo kombiniranje. Fleksijski generator proizvodi konačne oblike kao i paradigmu uzimajući u obzir fleksijski uzorak za hrvatski jezik kao što je definirano i detaljno opisano u Tadić (1994)[48] (Tadić-Fulgosi 2003: 42).

[46] Hrvatski je morfološki leksikon dostupan na internetskoj stranici http://hml.ffzg.hr/hml/.

[47] Prijevod MA. Izvornik:„The very idea of generating all possible combinations of morphemes appeared in Halle (1973) for the first time in the framework of generative grammar. Although at that time it seemed that morphology could be represented as a compact subcomponent of the lexicon in the generative grammar, that concept posed several theoretical problems. (...) Our idea of modeling the morphology of Croatian was somewhere along that track but we wanted to keep the model as simple as possible and use the computational data from the research already completed for Croatian“.

[48] Prijevod MA. Izvornik: „Inflection: modeled as a list of generated stems and a list of inflectional morphemes with rules for their combining. The inflectional generator produces the final wordforms along the paradigms and with regard to inflectional patterns for Croatian as defined and described in detail in Tadić (1994).“

Međutim, u ovome se slučaju računalna morfološka analiza shvaća u vidu opisa Roberta Wołosza (2005) koji je u svome radu pisao o samome procesu računalne morfološke analize na primjeru poljskoga jezika te ju definirao na sljedeći način:

> Kao što vidimo, morfološka se analiza mora shvatiti u svome punom obliku: bit je u određivanju vrijednosti rječničkoga oblika leksema odnosno gramatičke kategorije, a ne leksičkih i gramatičkih morfema na koje se riječ može rastaviti. Dakle, s jezikoslovnoga stajališta ovo je dubinska morfološka analiza (vidi: Melcsuk 1974, Świdziński 1997), a ne površinska, koja se često rabii na različitim razinama radova didaktičke prirode. Kod efektivne morfološke analize riječi dijelim na dijelove, ali ova je podjela isključivo tehničke prirode, te se zbog toga ne koristim jezikoslovnom terminologijom pri njihovu imenovanju, poput primjerice korijena, nastavka, morfema itd.
> Pod pojmom morfološka sinteze (generiranje) podrazumijevam proces koji je obrnut od analize, odnosno stvaranje oblika riječi koje pripadaju danom leksemu i odgovaraju željenoj gramatičkoj kategoriji. Ako dakle primjerice želim generirati akuzativ jednine ženskoga leksema *noga*, onda rezultat morfološke sinteze mora biti oblik *nogę*[49] (Wołosz 2005: 90-91).

Detaljniji opis procesa morfološke analize dan je u poglavlju 4.1.

Kao što je iz analize udžbenika koji se rabe u nastavi hrvatskoga jezika na njemačkome govornom području i jezičnih priručnika koji se rabe općenito u nastavi hrvatskoga jezika vidljivo, potrebno je bilo razviti rječnik koji bi se temeljio na stvarnoj jezičnoj uporabi. Procesom se računalne sinteze, kao što je i Tadić (1994: 31) spomenuo, većinom provjeravaju istraživačke hipoteze, čime se podrazumijeva proizvođenje, a ne prepoznavanje oblika. Analizom elektronskoga jezičnog priručnika, Gramatičkoga tezaurusa, došlo se do zaključka da je GT rađen najvjerojatnije prema načelu morfološke sinteze. Razlozi koji upućuju na takav zaključak odnose se ponajprije na analizu pogrješaka, a koje se mogu protumačiti kao rezultat korištenja procesa generiranja oblika.

Primjerice, kao jedna od pogrješaka koje su nastale najvjerojatnije korištenjem toga procesa jest činjenica da je u GT primjerice kod imenica *talog, prilog, zalog* (koje pripadaju kategoriji *n082* u GT) navedeno da one nemaju množinu, što je u raskoraku s navodima u Anićevu rječniku. Nadalje, kod mnogih imenica nisu navedeni dvojni leksemi u množini, primjerice kod imenice *san*, gdje se kao jedina množina pojavljuje paradigma s leksemom *snovi* u nominativu množine, izostavljajući kao mogući oblik i paradigmu s leksemom *sni* u istome padežu i broju, a koja se navodi u Anićevu rječniku. Pogrješke kod navođenja rodova između ostaloga isto tako upućuju na sintezu koja se vjerojatno temeljila

[49] Prijevod s poljskoga na mađarski – Robert Wołosz, s mađarskoga na hrvatski – Melita Aleksa

na automatskoj obradi prema korijenima riječi i/ili afiksima. Rješenje za izbjegavanje ovakvih i sličnih pogrješaka leži u korištenju procesa automatske morfološke analize, gdje se leksički temelj programa mora u što većoj mjeri podudarati sa stvarnom jezičnom uporabom.

Pri izradi programa za morfološku analizu hrvatskoga jezika nastojala se postići što veća prepoznatljivost teksta i time što više približiti idealnome postotku od 100%. U leksički su se temelj morali uvrstiti svi oblici neke riječi, a ne samo oni koji su navedeni u referentnim izvorima (primjerice oblici već spomenute leme *san* koji nisu navedeni u RHJ). Time leksički temelj ne sadrži samo leme koje se nalaze u referentnim izvorima, nego je proširen lemama i oblicima koji se pojavljuju u testnome korpusu odnosno stvarnoj jezičnoj uporabi (više o samim detaljima izrade programa i o njegovome leksičkom temelju može se pronaći u poglavlju 4.2.).

Korištenjem leksičke baze analizatora koja je izrađena u svrhu morfološke analize za izradu morfološkoga rječnika, te gramatičkih kategorija stemova i termova koji se nameću uslijed njegova kreiranja, osigurala se izrada računalnoga morfološkog rječnika koji bi trebao u što većoj mjeri odražavati stvarnu jezičnu uporabu i na taj način pridonijeti što boljem usvajanju hrvatskoga kao drugoga ili stranoga jezika.

Logičnim se slijedom ovdje postavlja pitanje izabiranja jezične inačice za sastavljanje morfološkoga rječnika. Kratak prikaz te problematike prije opisa načela rada morfološkoga analizatora slijedi u sljedećem odlomku.

3.2. Izbor jezične inačice za izradu morfološkoga rječnika

Pri izradi morfološkoga rječnika koji će se temeljiti na procesu morfološke analize tekstova napisanih na hrvatskome jeziku, prvim se problemom pojavila dvojba oko izravnoga prenošenja građe morfološkoga analizatora na rječnik ili njezina reduciranja. Dvojba se odnosila na izbor samo jedne, pisane standardne inačice hrvatskoga jezika čiji bi temelj bila građa iz Anićeva rječnika, ili inkorporiranje i onih sastavnica rječničke baze koje su dodane osnovici analizatora, odnosno Anićevu rječniku (više o tome u poglavlju 6.2.). U ovom će se poglavlju ukratko predočiti problematika koja se pojavila pri izradi

morfološkoga rječnika, a koja je povezana s dvojbama koje su se pojavile pri izradi analizatora, koje su pojašnjene u poglavlju 8.

Naime, izbor jezične inačice koja će činiti srž morfološkoga rječnika, analogno problematici izbora jezične inačice kod razvijanja analizatora, sa sobom prvenstveno donosi problem izbora pisanoga ili govornoga jezika. Problemi se nadalje odnose na činjenicu treba li u sastav morfološkoga rječnika uključiti sastavnice koje su prisutne u inačicama hrvatskoga jezika.

Ako se promotri analiza računalnih rječnika opisana u poglavlju 9.1.1., zaključuje se da bi rječnik trebao pokrivati što više sastavnica koje mogu poslužiti učenju nekoga jezika. Time se najprije postavlja pitanje uvođenja sastavnica razgovornoga jezika u temelj morfološkoga rječnika. Pri tome se pojavio problem kodifikacije razgovornoga jezika, odnosno nedostatak izvora kojim bi se pokrile prozodijske karakteristike svih oblika neke leme. Naime, unatoč tomu što Anić (2000) u svome rječniku navodi naglaske na lemama i pojedinim oblicima kojima se koristi u rječniku, pojavljuje se problem cjelovitosti određivanja prozodijskih karakteristika ostalih oblika riječi, odnosno problem referentnih izvora gdje bi bili navedeni svi oblici svih riječi u rječniku sa svojim naglasnim oznakama (više o tome u poglavlju 6.2.). Iako je prema Silić-Pranjkoviću (2005: 42) naglasak sastavni dio osnove, pa i on može sudjelovati u njezinoj promjeni, u ovome se slučaju morfološka analiza i paradigme u morfološkome rječniku temelje na računalnim morfološkim kategorijama te je kategorija naglaska u ovoj fazi projekta izostavljena. Daljnjim se problemom prikazala i cjelovitost morfološkoga opisa riječi jer se upotrebom prozodijskih sastavnica broj stemova i termova uvelike povećava (više o tome u poglavlju 6.2.).

Zbog navedenih se razloga, kao prvotnim izborom, pojavilo ograničenje rječnika na pisani jezik, s mogućnošću uključivanja i oznaka naglasaka u nekom od sljedećih projekata, odnosno pri izradi programa za prepoznavanje ljudskoga govora. Tada bi se proširio leksički temelj programa.

Sljedeće pitanje koje se postavilo pri izradi rječnika jest pitanje izbora jezične inačice, što se prvenstveno odnosilo na problematiku uključivanja sastavnica ostalih inačica (regionalnih, dijalektalnih, itd.) u leksički temelj rječnika. Naime, u *Kurikulumu* su ove sastavnice uvrštene jer se između ostaloga navodi:

> Hrvatski standardni jezik temelj je nastave hrvatskoga jezika u inozemstvu, ali treba iskoristiti i učeničku imanentnu gramatiku - gramatički sustav zavičajnog dijalekta.
> (Kurikulum hrvatske nastave u inozemstvu,
> http://public.mzos.hr/Default.aspx?sec=2116, 4. lipnja 2008.)

Iz ovoga se citata zaključuje da bi rječnik trebao sadržavati i sastavnice iz regionalnih inačica hrvatskoga jezika. Postupak kojim se neka lema uvrštava u sastav analizatora i morfološkoga rječnika sastoji se od nekoliko koraka. U prvome koraku, nakon analize testnoga korpusa, ako se ispostavi da je sastavnica često zastupljena, prelazi se na njezino potvrđivanje. Ono se provodi analizom čestotnosti njezina pojavljivanja u cjelokupnome korpusu i uspoređivanjem čestotnosti s HČR. Ako se naime analizom testnoga korpusa pronađu leksemi ili oblici koji su česte pojavnice, a koji nisu dijelom hrvatskoga standardnoga pisanog jezika, odnosno nisu zastupljeni u RHJ, oni se uvrštavaju u leksički temelj analizatora. Iz razloga što se morfološki rječnik izrađuje za učenje hrvatskoga standardnog jezika, sastavnice za koje se utvrdi da nisu u rječničkome sastavu hrvatskoga jezika, ali su primjerice često zastupljene u testnome korpusu (primjerice oblik *bodibilder* koji je često zastupljen u internetskome potkorpusu), ipak neće biti uvrštene u sastav morfološkoga rječnika (više o tome u poglavlju 8.) namijenjenog učenju i poučavanju hrvatskoga standardnog jezika.

Iako je analizom udžbenika za učenje hrvatskoga kao drugoga ili stranoga jezika na njemačkome govornom području utvrđeno da se u korpusu pojavljuju regionalne inačice (poglavlje 2.2.), pri izradi morfološkoga analizatora došlo se do zaključka da je količina materijala napisana na inačicama hrvatskoga jezika ograničena i prilično nedostupna. Time se došlo do zaključka da bi se pri izradi rječnika ipak trebalo odlučiti za standardni jezik, s mogućnošću proširenja rječničkoga temelja jednoga dana onim čestim pojavnicama koje nisu u sastavu RHJ, a koje pripadaju ostalim inačicama hrvatskoga jezika. Ovoj činjenici ide u prilog činjenica da u sastav *Kurikuluma nastave hrvatskoga jezika u inozemstvu* nije uvršteno detaljnije usvajanje ostalih inačica hrvatskoga jezika, nego samo razvijanje receptivnih vještina. To se između ostaloga potkrjepljuje sljedećim navodima[50]:

> **razlikovati** hrvatski standardni jezik i hrvatska narječja (…)

[50] *Kurikulum nastave hrvatskoga jezika u inozemstvu* dostupan je na internetskoj stranici http://public.mzos.hr/Default.aspx?sec=2116 (4. lipnja 2008.).

> **prepoznavati** razlike u hrvatskom standardnom jeziku i hrvatskim narječjima (štokavskom, čakavskom, kajkavskom; tekstovni i glazbeni primjeri, zavičajne skladbe, uspavanke) (…)
>
> upotrijebiti odgovarajuća jezična sredstava u skladu s određenom govornom situacijom, **utvrditi razlike** hrvatskoga standardnoga jezika i hrvatskih narječja (…)
>
> pravilno izgovarati glasove hrvatskoga standardnoga jezika i **uočiti razlike** u odnosu na izgovor, slušanje glasova u hrvatskim narječjima (prema jezičnom iskustvu učenika) (…)
>
> **istražiti** na primjerima razlike u hrvatskom standardnom jeziku i hrvatskim dijalektima (hrvatsko dijalektalno pjesništvo)[51].(Kurikulum hrvatske nastave u inozemstvu, http://public.mzos.hr/Default.aspx?sec=2116, 4. lipnja 2008.)

Detaljniji opis procesa morfološke analize na primjeru prilagodbe morfološkoga analizatora, programa HUMOR-a hrvatskom jeziku, njegovih karakteristika, problematika izrade korpusa, jezične politike te izabiranje jezične inačice koja će činiti osnovicu leksičke baze morfološkoga analizatora prikazani su u sljedećem poglavlju.

[51] Isticanja moja.

4. HUMOR kao jedno od sredstava za računalnu morfološku analizu

HUMOR, odnosno High–Speed Unification based Morphology, jest morfološki analizator (parser) koji se temelji na unifikaciji koji je razvila tvrtka MorphoLogic. Danas se morfološki parsing ili morfološka analiza rabi na mnogim područjima, prije svega za analizu različitih jezika, a nadalje i kao temelj za mnoge računalne programe. Razlog odabiranja HUMOR-a kao sredstva za morfološku analizu hrvatskoga standardnog jezika leži u praktičnosti, odnosno projektu koji je započeo na iniciranje začetnika tvrtke MorphoLogic, a odnosio se na prilagođavanje sustava HUMOR-a sustavu hrvatskoga standardnog jezika. Kao drugi važniji razlog jest široka primjena programa, jer je on danas postao osnovica za prevoditeljski softver koji je u širokoj primjeni, a koji je razvijen u suradnji s Mađarskom akademijom znanosti.

Kao što i sami autori navode, HUMOR je program s karakteristikama univerzalnosti, što podrazumijeva njegovu implementaciju za različite jezične sustave te modificiranje u svrhu razvijanja različitih računalnih pomagala. Iznimna brzina provođenja procesa analize (otprilike se 25 000 riječi/s analizira na prosječnom računalu) glavna je karakteristika HUMOR-a, kao i činjenica da se radi i o djelomičnome samoispravljačkom sustavu. To podrazumijeva činjenicu da se testiranjem na jednome korpusu ukazuje na pogrješke koje su kasnije podložne ispravljanju i poboljšavanju. Kada je riječ o načinu rada samoga HUMOR-a, važno je spomenuti činjenicu da univerzalnost programa dokazuje i to da je pomagalo razvijeno u početku za morfološku analizu aglutinativnih jezika (prvenstveno mađarskoga jezika), ali da su u međuvremenu dorađene i druge verzije fleksijskih jezika (engleski i njemački), visokofleksijskih jezika (poljski i ruski), dok se trenutno provodi testiranje većine europskih jezika, među kojima je i staroslavenski. Univerzalnost programa osim već navedene mogućnosti prilagodbe različitim jezičnim sustavima dokazuje i činjenica da se on implementirao i u različita pomagala, primjerice MobiMouse (jezično pomagalo koje se rabi za prevođenje riječi i izraza koji se nalaze na korisnikovu ekranu), MobiDic (dvosmjerni računalni rječnik) i MetaMorpho (prevoditeljski sustav koji se rabi za prevođenje jednostavnih rečenica s engleskog i na engleski jezik), čija je najnovija inačica MetaMorphoTermX, pomagalo koje između ostaloga sažima uneseni jezični materijal, popisuje pojavnice s obzirom na njihovu učestalost pojavljivanja u

tekstovima, analizira ih u pogledu konkordancije te daje njihov prijevod (više o navedenim pomagalima u poglavlju 9.2.).
Jedna od početnih ideja rada na hrvatskome HUMOR-u bila je razvijanje sličnih softvera koji bi se mogli koristiti i u nastavi hrvatskoga kao stranog jezika s hrvatskim kao polazišnim odnosno ciljnim jezikom. Iz razloga što format leksičkog i gramatičkog temelja računalnoga programa nije ovisan o HUMORu, nego se daje u općeprihvaćenom formatu, on se može rabiti u različite svrhe, a ne samo za morfološku analizu. Upravo se iz toga razloga i pojavila ideja o sastavljanju morfološkoga rječnika.

Iz razloga što je HUMOR bio polazište za morfološku analizu, sastavljanje morfološkoga rječnika u prvome i prevoditeljskih softvera u drugome redu, smatram bitnim u ovoj fazi rada, a u svrhu otklanjanja dvojbi koje bi se eventualno kasnije pojavile, detaljno predočiti proces morfološkoga parsinga i sastavnica koje se ovdje pojavljuju. Proces će biti predstavljen s jezikoslovne strane, a ne sa stajališta programiranja, usredotočit će se na jezičnu problematiku pri opisu programa, dajući samo prijeko potrebne tehničke informacije. Predstavit će se isto tako sam proces prilagodbe HUMOR-a jednom visokofleksijskome jeziku, te problemi i različita rješenja koja će se uspješno primijeniti i kod podučavanja hrvatskoga kao drugoga ili stranoga jezika. Osim navedenih problema prikazat će se i problematika te način prilagodbe HUMOR-a sustavima drugih jezika, s posebnim osvrtom na rješenja pri računalnoj morfološkoj analizi njemačkoga, mađarskoga i hrvatskoga jezika. Iz razloga što navedeni problemi nisu samo jezične prirode, nego se izravno odnose i na problematiku jezične politike, prikazat će se i rješenja koja su primijenjena pri prilagodbi HUMOR-a sustavima drugih slavenskih jezika. Važno je odmah napomenuti da se HUMOR kao sustav u ovome radu koristi samo kao primjer za morfološku analizu, odnosno za bolje pojašnjavanje računalno-lingvističkih pojmova, a ne u reklamne svrhe za sam program.

4.1. Način rada HUMOR-a

Prva je demo verzija morfološkoga parsera HUMOR-a koji se temelji na unifikaciji nastala 1992. godine s ciljem morfološke analize jezika i njegove implementacije u različite jezične alate. Glavnim se ciljem nije postavio razvoj programa za rastavljanje riječi na

slogove (tzv. hyphenatora), programa za provjeravanje pravopisa i tezaurusa jer su takvi sustavi već duže vrijeme prisutni na tržištu, nego jezikoslovni parsing lema kao i dubinska i površinska analiza za potrebe pretraživanja u sustavima koji podupiru prevođenje (Proszéky–Kis 1999a: 266). Korištenje HUMOR-a pojasnili su i sami autori, Prószéky–Kis (1999aa), koji navode da

> primarna svrha korištenja morfološkoga analizatora leži u cilju da se dobije što je više moguće jezikoslovnih informacija o pojedinom obliku riječi koliko je moguće. Drugi cilj leži u korištenju osnovnih načela morfološke analize da bi se parser implementirao. To znači da možemo ili prikupiti ili generirati uzorke fraza na različitim lingvističkim razinama (imeničke, prijedložne, glagolske fraze) i sastaviti jedan leksikon svojstven Humor-u. Na posebnoj razini svaka pojedina sastavnica nekoga uzorka odgovara cjelovitijoj strukturi na nižoj jezikoslovnoj razini (Prószéky–Kis 1999a: 267)[52].

Kao što se iz ulomka može pročitati, jedna je od glavnih svrha korištenja sustava sama morfološka analiza tekstova, odnosno sastavljanje leksičkog temelja programa koji se dalje može koristiti u različite svrhe. Jednom se od njih u ovome radu pokazalo i sastavljanje morfološkoga rječnika koji se temelji na morfološkoj analizi (više o tome u poglavlju 8.).

Da bi se podrobnije objasnio način rada programa, moraju se pojasniti određeni pojmovi koji se tiču samoga procesa morfološke analize. Za razliku od tradicionalne analize, glavno načelo računalne morfološke analize leži u podjeli lema na *stemove* i *termove*. U ovome se slučaju namjerno ne koriste tradicionalne gramatičke kategorije korijena riječi i afiksa, jer se one ne podudaraju uvijek s navedenim pojmovima. Ukratko pojašnjeno, kategorija stemova, naime, uključuje one dijelove riječi koji kroz cijelu paradigmu ostaju nepromijenjene, dok se preostali dio leme naziva termom. Važno je pri tome spomenuti i činjenicu da se slično kao u tradicionalnoj morfologiji riječ ne može sastojati od Ø–stema, ali može imati Ø–term. Primjerice ako pomnije promotrimo pridjev *hrvatski*, možemo uočiti da se u ovome slučaju stemom označuje dio riječi *hrvatsk–*, dok je term u ovome obliku *–i.* Analogno se tomu da zaključiti da će u leksičkoj bazi parsera kod pojedinih lema jedan stem biti povezan s velikim brojem termova s kojima se on može kombinirati da bi

[52] Prijevod MA: „the first point of using the morphological analyzer in the parser is to get as much linguistic information about a single word form as possible. The second point is using the basic principles of the morphological analyzer to implement the parser itself. This means that we either collect or generate phrase patterns on different linguistic levels (noun phrases, prepositional phrases, verbal phrases etc.) and compile a Humor–like lexicon of them. On a specific linguistic level, each atomic element of a pattern actually corresponds to a (more) complex structure on a lower linguistic level".

se tvorili različiti oblici. Stemu *hrvatsk–* u ovome se slučaju dodaju primjerice termovi *–a, –oga* i *–ih,* da bismo u paradigmi dobili odnosno da bi analizator kasnije prepoznao oblike *hrvatska, hrvatskoga, hrvatskih*...

Leksikon HUMOR-a ne sadrži samo jedan stem za svaku pojedinu lemu, nego alomorfe stemova, dok se termovi nalaze u posebnoj ulaznoj jedinici zajedno s kodovima pomoću kojih se oni pridružuju svojim stemovima. Time dobivamo ograničen broj termova koji se mogu pridružiti stemovima, odnosno ograničen broj krajnjih oblika riječi koji su dijelovi paradigme. U slučaju engleskoga jezika autori to pojašnjavaju na sljedeći način:

> Leksikon Humora 99[53] sadrži alomorfe stemova (koji su generirani tijekom faze učenja koja je prije spomenuta), a ne pojedinačne stemove. Odnosi između alomorfa istog korijena (primjerice *wolf, wolv*), međutim, važni za sintaksu, semantiku i krajnjega korisnika. Jedan online morfološki analizator ne mora se izravno baviti derivacijom alomorfa od njihova osnovnog oblika, primjerice kako je *happi* nastao od *happy* ako je ispred *–ly*. Ovaj fenomen – posljedica ortografskoga sustava – rješava se lingvističkim procesom Humora 99 offline, što analizu čini mnogo bržom. Ova je metoda slična metodi kompiliranja leksikona u finitnim modelima[54] (Prószéky–Kis 1999a: 262).

U slučaju prilagodbe sustavu hrvatskoga jezika, alomorfi stemova podrazumijevaju primjerice stemove koji su potrebni za stupnjevanje pridjeva jer se time uvelike ubrzava morfološka analiza i preskače se programiranje u kojemu bi program „učio“ kako primjerice stvoriti stem za stupnjevanje pridjeva od temeljnoga oblika riječi. Univerzalnost programa naime omogućuje proširivanje leksikona proizvoljnim brojem novih lema i stemova, što omogućuje i bolji postotak prepoznavanja tekstova te povećava pouzdanost programa.

Za vrijeme morfološke analize pod površinom se događa djelomična odnosno cijela analiza (vidi: Prószéky–Kis 1999a), zbog čega se krajnji rezultat uvelike razlikuje od rezultata na potpovršinskoj razini. U slučaju mađarskoga jezika autori su dva procesa prikazali na primjeru oblika glagola *elszámítógépezgethettem* ('mogao sam se služiti računalom neko vrijeme da se zabavim'):

[53] Humor 99 radni je naziv programa HUMOR-a, koji se rabio tijekom njegove izrade.

[54] Prijevod MA: „Humor 99 lexicons contain stem allomorphs (generated by the learning phase mentioned above) instead of single stems. Relations among allomorphs of the same base form (e.g. wolf, wolv) are, however, important for syntax, semantics, and the end–user. An online morphological parser needs not be directly concerned with the derivation of allomorphs from their base forms, for example, it does not matter how happi is derived from happy before –ly. This phenomenon – a consequence of the orthographical system – is handled by the off–line linguistic process of Humor 99, which makes the analysis much faster. This method is close to the lexicon compilation used in finite–state models.“

INPUT:
elszámítógépezgethettem
INTERNAL SEGMENTATION:
el[PREFIX]+számító[STEM 1]+gép[STEM2]+ezgethet[DERIV.AFF.]+tem[INFL.AFF]

OUTPUT:
el[VPREF]+számító[ADJ]+gép[N]+ez[N2V]+get[FREQ]+het[OPT]+tem[PAST–SG– 1]
(Prószéky–Kis 1999a: 266)

Output na površinskoj razini sadrži tradicionalne lingvističke oznake i kategorije koje se mogu dalje koristiti. U ovome se primjeru naime radi o površinskoj analizi koja je specifična za aglutinativne jezike te se razlikuje od analize fleksijskih jezika kao što je vidljivo iz primjera (1) i (2).
Prednosti ovakvog načina obrade podataka opisani su u sljedećem odlomku:

> Dvorazinska je morfologija reverzibilni sustav koji se temelji na ortografiji, a koja ima nekoliko prednosti s jezikoslovnoga stajališta. Naime, morfo-fonetska / grafemska pravila mogu se formalizirati na općeniti i vrlo elegantan način. Ona isto tako sadrži i računalno-lingvističke prednosti, ali leksikon mora sadržavati ulaze s posebnim simbolima i drugim sofisticiranim elementima da bi se stvorili potrebni površinski oblici. Korisnicima koji nisu jezikoslovci treba rječnik koji mogu lako proširivati, u koji se riječi mogu unositi (gotovo) automatski. Leksički temelj Humora 99 sadrži površinske znakove – bez transformacija – dok metarječnički mehanizmi zadržavaju mnoge prednosti dvorazinskoga sustava. U praksi to znači da korisnici mogu dodavati rječničke unose u sustav koji je pokrenut bez njegova ponovnoga sastavljanja[55] (Prószéky–Kis 1999a: 266).

Iz razloga što je cilj ovoga rada opis procesa morfološke analize koji se rabi u svrhu razvijanja programa za učenje i poučavanje hrvatskoga kao drugoga ili stranoga jezika, neće se podrobnije ulaziti u proces potpovršinske analize, nego će se podrobnije promotriti površinska analiza s jezikoslovnoga stajališta.

Na temelju unesenih informacija i morfoloških karakteristika termova, pri analizi pojedinih riječi, oblik se riječi analizira kako je prikazano u primjerima (1) i (2). U ovome se slučaju

[55] Prijevod MA: „Two–level morphology is a reversible, orthography–based system that has several advantages from a linguist's point of view. Namely, the morpho–phonemic/ graphemic rules can be formalized in a general and very elegant way. It also has computational advantages, but the lexicons must contain entries with extra symbols and other sophisticated elements in order to produce the necessary surface forms. Non–linguist users need an easyto– extend dictionary into which words can be inserted (almost) automatically. The lexical basis of Humor 99 contains surface characters only – no transformations are applied –, while the metadictionary mechanism retains many advantages of the two–level systems. It means in the practice that users can add entries to the running system without re–compiling it.“

radi o obliku *stranke* koji smo zadali programu za analizu, a prikazana je površinska analiza, odnosno output.

Analysis of "stranke"
(1) stranka[Sf]=stran+ke[21]
(2) stranka[Sf]=stran+ke[12;42;52]

Navedeni primjer pokazuje da je program zadani oblik povezao s više pojavnica leme *stranka*. Prva je riječ u izlaznome rezultatu uvijek leksem. Oznaka iza leksema jest oznaka vrste riječi i roda (u ovome se slučaju radi o imenici ženskoga roda), leme *stranke* koja se spominje u nastavku u obliku stema i termova. Stemom se u ovome primjeru smatra oblik *stran–*, jer taj dio riječi ostaje nepromijenjen u cijeloj paradigmi, dok je *–ke* u ovome slučaju term koji se nadovezuje na stem i odvojen je od njega znakom +. Prvi broj u zagradi označuju padež, dok se drugi broj odnosi na broj.
Kada se ovi rezultati interpretiraju, može se uočiti da je analizator leksem *stranke* povezao s različitim oblicima leme *stranka.* Moguće su interpretacije oblika u genitivu jednine [21], nominativu množine [12], akuzativu množine [42] i vokativu množine [52].
U pojedinim se slučajevima određeni oblik može asocirati i s različitim lemama, kao što je vidljivo u sljedećem primjeru:

Analysis of "damo"
(3) dama[Sf]=dam+o[51]
(4) dati[Vs]=da+mo[ip1]
(5) dati[Vs]=da+mo[p1]

Damo se kao oblik riječi, ako se izuzme kontekst, može interpretirati kao imenička riječ u vokativu jednine od imenice *dama*, kao glagol u prvom licu množine imperfekta glagola *dati* [ip1] ili kao prvo lice množine prezenta glagola *dati* [p1]. Kada se ovakva vrsta analize unese u neki prevoditeljski softver, važno je uzeti u obzir činjenicu da bi program trebao ponuditi prijevode obaju lema, i imenice *dama* i glagola *dati* i na ciljnome jeziku. Ovisno o rezultatima potpovršinske analize oblici riječi koji se pojavljuju u polazišnome jeziku povezuju se nadalje s lemama u ciljnome jeziku.

Pri objašnjavanju morfološke analize Wołosz (2005) razlikuje više kategorija: *słowo* (riječ), *forma hasłowa* (rječnički oblik ili natuknica), *forma wyrazowa* (oblik riječi ili obličnica), *leksem* (leksem ili različnica). Najvažnijom se sa stajališta razvijanja morfološkoga analizatora za sustav hrvatskoga standardnog jezika pokazala definicija *forme wyrazowe* (hrvatski izraz: *obličnica*), koji ima višestruku vrijednost. Pod pojmom *obličnica* prvenstveno se podrazumijeva grafička razina riječi kojoj se pridodaje i određena gramatička vrijednost (primjerice *more,* nominativ jednine), te se time govori o jednome značenju, dok se u drugome redu grafičkoj razini pridodaje druga gramatička vrijednost (primjerice *more,* akuzativ jednine) (Jelaska 2005a: 139). Pri razvijanju gramatičke osnovice HUMOR-a i HumorEsk-a ove su se kategorije pokazale izrazito bitnima, kao i pri sastavljanju morfološkoga rječnika (više o tome u poglavlju 8.).

Primjeri (3), (4) i (5) isto se tako mogu prikazati poput tipičnoga problema stemminga. Proces je stemminga naime jezikoslovni proces koji se može smatrati funkcijom normaliziranja, koja 'normalizira' oblike riječi i sustavima za pretraživanje omogućuje pronalaženje pojedinoga oblika riječi bez obzira na oblik koji je unesen u tražilicu (Prószéky–Kis 1999a: 267). Kod jezika u kojima jedna lema može poprimiti više oblika, potreban je ovakav sustav normalizacije, posebice kod sustava za prevođenje te elektronskih rječnika (Prószéky–Kis 1999a: 267). U slučaju kao u primjerima (3), (4) i (5), prema riječima tvoraca morfološkoga analizatora, krajnja odluka o odabiru odgovarajućega oblika leži na samome korisniku ili na kontekstnome disambiguatoru (Prószéky – Kis 1999a: 267). Ova činjenica ide u prilog spoznaji da računalo još uvijek ne može u potpunosti obaviti sve one zadaće koje obavlja čovjek.

4.2. Struktura morfološkoga analizatora

Uzimajući specifičnosti svake jezične skupine u obzir, struktura se HUMOR-a mora prilagoditi sustavu svakih od pojedinih jezika. U ovome će se dijelu rada pojasniti struktura programa te razlike između strukture parsera za aglutinativne i za fleksijske jezike, s posebnim naglaskom na slavenske inačice parsera – hrvatsku i poljsku varijantu. Dat će se prikaz baza podataka koje su dijelovi HUMOR-a s jezikoslovnoga stajališta. Tehnička će se izvedba programa u ovome poglavlju spomenuti samo u svrhu boljega pojašnjavanja

načela rada samoga parsera, ali se neće dodatno objasniti proces morfološkoga parsinga. Razlog tomu leži u preuzimanju strukture morfološkoga analizatora i u postavljenim lingvističkim ciljevima ovoga rada. Prikaz načina rada parsera kod analize aglutinativnih jezika dat će se iz razloga što se želi dokazati univerzalnost samoga programa te njegova prilagodljivost različitim jezičnim sustavima.

U osnovici se morfološki parser HUMOR sastoji od dvaju velikih dijelova. Osim tehničkoga dijela programa, tzv. *engine*-a, koji se može okarakterizirati kao pogonski dio samoga parsera, postoje najmanje dvije velike baze podataka pomoću kojih se događa morfološka analiza. Način njihova strukturiranja pojašnjava Novák (2003) u svome radu gdje navodi:

> Preduvjet uspješnoga rada programa jest da ispitivanja koja se događaju za vrijeme analize budu vrlo jednostavni i brzi procesi. Tomu je potrebno da baze podataka sadrže veliku količinu zalihosnih informacija u eksplicitnom obliku, da se one ne bi morale izračunavati za vrijeme analize[56] (Novák 2003: 2).

Baze se podataka mogu proširivati uslijed uzimanja u obzir specifičnosti pojedinih jezika te praktičnosti same izvedbe programa, o čemu će biti više govora u sljedećem odlomku. Jedna se baza podataka može nazvati leksičkom bazom podataka, jer sadrži STEM-ove, dok druga baza podataka sadrži TERM-ove s pripadajućim kodovima i uzorcima tvorbe koji su sadržani u matrici programa. Dalje će se u radu koristiti izrazi *leksička osnovica* programa za bazu podataka sa stemovima i *gramatička osnovica* programa za bazu podataka s termovima.

Leksička se osnovica programa naime sastoji od rječničke baze koja je preuzeta iz jednojezičnih rječnika jezika. U slučaju hrvatskoga jezika ona trenutno sadrži cijeli rječnički opus Anićeva rječnika. Razlozi odabiranja Anićeva rječnika leže u činjenici da je projekt započeo u vremenu kada nisu bili dostupni drugi rječnici hrvatskoga jezika (poput primjerice Šonjina (2000) rječnika). Kako je već prije navedeno, leksička je osnovica samoga programa fleksibilan sustav koji omogućuje proširivanje natuknicama koje nisu u sastavu RHJ, a za koje se korpusnom analizom ispostavi da su česte pojavnice u hrvatskome korpusu (više o tome u poglavlju 6.3.). U hrvatskoj inačici morfološkoga

[56] Prijevod MA. Izvornik. „A program hatékony működésének az a feltétele, hogy az elemzés közben végrehajtandó ellenőrzések nagyon egyszerű és gyors műveletek legyenek. Ehhez az kell, hogy az adatbázis rengeteg redundáns információt tartalmazzon explicit formában, hogy ezeket ne elemzés közben kelljen kiszámítani.“

analizatora postoji trenutno 54 000 natuknica. Iako je u opisu samoga rječnika navedeno da RHJ sadržava 60000 lema, pri sastavljanju leksičke osnovice programa ispostavilo se da je broj različitih lema približno 54 000. U RHJ su naime neki rječnički unosi dvostruko navedeni, primjerice određeni i neodređeni oblici pridjeva, poput leme *crven* i *crveni*. U leksičkoj su osnovici parsera ovi oblici svrstani pod istu natuknicu, te je jedan oblik uzet za lemu, dok je drugi oblik naveden u paradigmi određene natuknice (više o pridjevskoj paradigmi iz spomenutoga primjera u poglavlju 7.3.).

Gramatička je osnovica programa zanimljivija sa stajališta izrade računalnoga morfološkog rječnika hrvatskoga standardnog jezika. Ona sadrži naime termove s pripadajućim kodovima, što zajedno sa stemovima za svaku pojedinu riječ čini paradigmu neke riječi. Iz razloga što se ovdje radi o računalnoj morfološkoj analizi, pri izradi gramatičke osnovice programa došlo se do rješenja koja se mogu uspješno primijeniti u nastavi hrvatskoga kao drugoga ili stranoga jezika jer svojim formalističkim pristupom mogu olakšati usvajanje gradiva, ali prije svega dati odgovor na konkretna pitanja učenika (više o tome u poglavlju 8.3.). Izradom termova za svaki pojedini stem u leksičkoj osnovici dobiva se osnovica računalnoga morfološkog rječnika za hrvatski jezik koja će sadržavati paradigme svih natuknica iz Anićeva rječnika.

Kao što se na početku ovoga odlomka navelo, osnovice su programa specifične za pojedine jezike. Za slavenske se jezike izabrala struktura spomenute leksičke osnovice i gramatičke osnovice parsera, dok se kod aglutinativnih jezika (mađarskoga) moralo primijeniti drugačije rješenje. Iz razloga što u mađarskome jeziku afikse ne može dobiti samo korijen riječi, nego i oblik riječi s korijenom i prethodno dobivenim afiksima, broj bi se termova znatno povećao te time približio granici nemogućnosti generiranja svih oblika. Iz toga se razloga pribjeglo stvaranju dodatne baze podataka koja se može zamisliti kao baza podataka koja se nalazi između leksičke i gramatičke baze, a naziva se baza podataka s derivatima, odnosno afiksima. Oni se mogu spojiti s lijevopozicioniranim stemom ili desnopozicioniranim termom. Proces se morfološke analize mađarskoga jezika s obzirom na korištene baze podataka može pojasniti na sljedeći način:

> Cilj analize riječi jest da se odredi od kojega korijena i uz pomoć kojih sufiksa (ili prefiska), odnosno kojom je vrstom tvorbe riječi nastala zadana riječ (npr. *legelemibb*: *leg-elem-i-bb*). U pojedinim jezicima, poput primjerice u engleskom, gdje se u tolikoj mjeri ne koriste prefiksi i sufiksi, analiza riječi može se riješiti u pravilu izradom samoga rječnika. U mađarskom jeziku to nije slučaj, jer sufiks se

> dodaje i na riječ koja je već dobila sufiks. U tome slučaju govorimo o relativnome korijenu. Redoslijed dodavanja sufiksa prilično je određen. Afiksi se mogu podijeliti na razrede, ovisno o tome na koje vrste riječi su primjenjivi i kojoj vrsti riječi pripada novostvorena rjieč. Različiti morfemi mogu imati isti oblik (npr. *kutyá-nak, lát-nak*). Određeni morfemi mogu promijeniti relativni korijen (varijante korijena, jednačenje).
> Ovisno o tome moraju se riješiti četiri zadatka:
> - Rezanje sufiksa (prefiksa)
> - „Vraćanje" promjena koje su se dogodile na korijenu nakon dodavanja afiksa
> - Provjera slaganja razreda riječi i afiksa
> - Određivanje vrste riječi korijena i uloge afiksa.
>
> Iako se ove četiri funkcije mogu razdvojiti, tijekom rješavanja zadatka one se isprepliću i jedna drugu nadopunjuju[57] (Farkas-Naszódi 1990: 22).

Pojašnjavajući gornji citat zaključujemo da tijekom morfološke analize mađarskoga jezika parser polazi od kraja riječi i „reže" slovo po slovo, sve dok ne dođe do stema riječi koji se nalazi u leksičkoj osnovici programa. Nakon toga slijedi grupiranje znakova, sve dok se ne dođe do prve skupine koja odgovara podatku iz gramatičke osnovice (terma) ili pak baze podataka s afiksima. Farkas-Naszódi (1990: 23), pojašnjavajući način pronalaženja korijena ili relativnoga korijena riječi, govore da se pri ovom procesu radi o jednome konačnom automatskom procesu, koji se sastoji od nekoliko koraka konačnoga broja, ali prije ili kasnije dođe se do krajnjega rješenja, odnosno prepoznavanja korijena riječi i pripadajućih mu morfema.

Iz razloga što slavenski analizatori ne moraju rješavati probleme koji su karakteristični za aglutinativne jezike, u slučaju hrvatskoga i poljskoga jezika parser radi na drugačiji način. Naime, tijekom morfološke analize stem se iz leksičke osnovice slaže s mogućim termovima koji mu pripadaju sve dok se ne identificira određeni oblik riječi. Međutim, program ne staje kod prvoga rješenja, nego nastavlja „potragu", sve dok nije provjerio sve

[57] Prijevod MA. Izvornik: „A szóelemzés célja, hogy megállapítsa azt, milyen szótőből és milyen toldalékolással (és előtagokkal), esetleg milyen szóösszetétellel jött létre az adott szó (pl. legelemibb: leg-elem-i-bb). Egyes nyelvekben, mint például az angolban, ahol a toldalékolás és előtagok alkalmazása szegényes, a szóelemzés gyakorlatilag egy szótár alkalmas kiépítésével megoldható. Nem így a magyarban, ahol a toldalékolt szó is kaphat toldalékot. (Ilyen esetekben relatív szótőről beszélhetünk.) A toldalékok szekvenciája eléggé kötött. A toldalékok (előtagok) is osztályokba sorolhatók aszerint, hogy milyen szóosztályokra alkalmazhatók, és milyen szóosztályba kerül az új szó. A különböző toldalékoknak lehet azonos az alakjuk (pl. kutyá-nak, lát-nak). Egyes toldalékok módosíthatják a relatív szótövet (tőváltozatok, hasonulások).Ennek alapján négy feladatot kell megoldani:
- A toldalékok (előtagok) levágása.
- A toldalékolás következtében végbement tőváltozások 'visszacsinálása'.
- A toldalékok és szóosztályok egyeztetésének ellenőrzése.
- A szótövek szófajának és a toldalékok szerepének meghatározása.

Bár a négy funkció szétválasztható, a megoldás során összefonódnak, egymást kiegészítik."

moguće termove koji pripadaju određenom stemu. Iz toga se razloga pojavljuje nekoliko rješenja te je potreban disambiguator za određivanje pojedinih natuknica.

Važno je pri tome napomenuti da se računalna morfološka analiza ograničava samo na područje riječi te se time interpunkcijski i specifični znakovi (poput primjerice crtice, zagrade, točke, dvotočja, trotočja, točke sa zarezom i razmaka) smatraju granicama analize. Prepoznavanje je složenih sintaktičkih struktura pri tome manjkavo, ali su se dosadašnja dostignuća pokazala dovoljnima za izradu računalnih prevoditeljskih sustava i dinamičnih rječnika koji uspješno prepoznaju i složenije sintaktičke strukture (više o tome u poglavlju 9.2.). O specifičnostima hrvatske inačice parsera bit će riječi u poglavlju 8.

Jedno od pitanja koje se postavlja u ovoj fazi rada odnosi se na pitanje izbora načela rada analizatora, koji se temelji na stemovima i termovima. Razlog tomu leži u činjenici da ako se preuzmu tradicionalne kategorije i pravila pri izradi paradigmi (primjerice pravilo kada određeni glas prelazi u neki drugi ili pravila o jednačenjima), postaje gotovo nemoguće računalno obraditi navedene paradigme i računalu „pojasniti“ kako tvoriti određene oblike, jer pri tome dolazi do mnogo oprečnih stvari.
Novák-Wensky (2007: 164) tako tvrde da u gramatikama koje su pisane „ručno“ postoji mnogo stvari koje ostaju u određenoj „magli“. Te se stvari u računalnoj gramatici moraju postaviti eksplicitnima, tako da oni jezični sustavi koji su primijenjeni u nekome računalnom sustavu postaju točniji od onih koji nisu iskušani preko računala. Sredstva za analizu i generiranje pružaju mnogo veće mogućnosti testiranja i pri tome su toliko temeljita, da se to na ručni način ne može ni zamisliti. „Iz razloga što stroj radi 'bez mozga', svaku pogrješku s kojom se susretne, prikaže bez pomilovanja[58]“ (Novák-Wensky 2007: 164).
Pri razvijanju svoga sustava TWOL-a, na kom se temelji HUMOR, Karttunen-Beasley (2001: 9) tako govore o povijesti razvoja morfološke analize, o pojedinim fazama te navode:

> Tijekom ovoga rada postalo je jasno da jezikoslovci teško ovladavaju *two-level* formalizmom. Mnogo je lakše pisati pravila koja su u sukobu jedna s drugima. Bilo

[58] Prijevod MA. Izvornik: „Mivel a gép „ész nélkül” dolgozik, minden hibát könyörtelenül kimutat, amivel szembetalálkozik.“

> je potrebno razviti i sustav njihove provjere i automatski odstranitti najvažnije primjere sukoba[59] (Karttunen-Beasley 2001: 9).

Zbog toga se pribjeglo formalističkome rješenju koje na taj način može pridonijeti uspješnijem usvajanju paradigmi, a time i učenju i poučavanju jezika. Rješenja nastala uslijed formalizma pri opisivanju jezične strukture kojom se koristi parser, dovela su do nekih saznanja koja se mogu rabiti i u nastavi hrvatskoga kao drugoga ili stranoga jezika. Pokusno istraživanje o samoj funkcionalnosti rječnika dalo bi pouzdanije rezultate te bi dokazalo ili pobilo navedenu hipotezu.

4.3. Definiranje pojmova

Prije samoga pojašnjenja načina rada morfološkoga parsera HUMOR-a i morfološkoga analizatora potrebno je definirati nekoliko pojmova i pojasniti terminologiju kojom će se koristiti u nastavku rada. U ovome poglavlju slijedi kratak pregled najvažnijih pojmova koji će spominjati te podrobnija analiza pojmova natuknutih u prethodnim poglavljima.

Prvi problem koji se pojavljuje i koji je potrebno pojasniti odnosi se na problem definiranja morfoloških kategorija te izuzimanje sintakse iz cjelokupne analize. Međutim, prije nego se prijeđe na pojašnjavanje karakteristika morfološke analize, potrebno je objasniti problematiku koja se izravno veže uz taj problem, a to su ograničenja procesa morfološkoga parsinga.

Naime, jedno od manjkavosti morfološke analize jest ograničavanje na razinu riječi, odnosno nemogućnost procesiranja složenijih struktura zbog razmaka kao granice analize. O toj je problematici pisao i Prószéky (2003) koji navodi:

> Ljudski čitatelj naime u nazivu Mađarska akademija znanosti vidi drugačiji razmak nego što su razmaci u rečenici u kojoj se taj naziv nalazi: ovaj razmak – umjesto da bi rastavio – upravo spaja riječi koje se nalaze oko njega. Za računalo, međutim, oba razmaka nose isti brojčani kod. Međutim, budimo iskreni. Pojedini programi za obradu teksta omogućuju unošenje tzv. nelomljivih razmaka, ali njihovo korištenje nije tipično za tekstove koji su zabilježeni na elektronički način, tako da se njima ne

[59] Prijevod MA. Izvornik: „In the course of this work, it soon became evident that the two-level formalism was difficult for the linguists to master. It is far too easy to write rules that are in conflict with one another. It was necessary to make the compiler check for, and automatically eliminate, most common types of conflicts."

> koristimo ni u pretraživačima. Kao jedina druga mogućnost ostaje jezični procesor, koji uz pomoć analizatora koji se proteže na više riječi i uz pomoć rječnika prepoznaje da pojedine riječi više pripadaju cjelini nego druge[60] (Prószéky 2003: 8).

Iz ovoga je citata vidljivo da pri morfološkome parsingu postoje određene granice analize. Sam parser HUMOR naime nije u mogućnosti s velikom točnošću odrediti koje riječi tvore jednu cjelinu te se time ograničava samo na razinu riječi. Korištenjem analizatora naziva HumorESK, kako je već prije pojašnjeno, koji sadrži i dodatna sintaktička znanja, olakšava se prepoznavanje primjerice složenijih glagolskih vremena što igra vrlo važnu ulogu pri izradi alata za prevođenje (više o tome u poglavlju 9.2.). Osim razmaka kao granice analize, prepreku HUMOR-u znače i interpunkcijski znakovi, poput primjerice točke, zareza, kose crte ili spojnice. Problem se pojavljuje pri obradi onih riječi gdje je znak sastavni dio same riječi, kao što je bilo navedeno u poglavlju 2.3.3. Naime, ako je kosa crta sastavni dio riječi poput kratice za motorni brod *m/b*, riječ se mora kao takva unijeti u sastav rječničke osnovice programa da bi se na odgovarajući način mogla prepoznati i analizirati.

Pri izradi HUMOR-a postavilo se time bitnim definirati određene kategorije te time razdvojiti morfološku razinu od sintaktičke razine. Kao prva kategorija koja se mora definirati jest kategorija riječi.

Riječ se naime u analizatoru percipira na grafički način, odnosno pod jednom se riječju smatra niz znakova koji se nalaze između graničnika – dvaju znakova razmaka ili interpunkcijskih znakova. Problem se kod ovoga načina percipiranja riječi pojavljuje kod polusloženica koje su odvojene spojnicom i kod složenih glagolskih vremena, odnosno kod upotrebe primjerice povratno-posvojne zamjenice.

Ako se naime pogleda definiranje uporabe spojnice u *Hrvatskom pravopisu* (Babić-Finka-Moguš 1995), vidimo da se daje sljedeća definicija:

[60] Prijevod MA. Izvornik: „A humán olvasó viszont tulajdonképpen másféle szóközt lát például a Magyar Tudományos Akadémia nevében, mint az azt befoglaló mondat más szavai között: ez a szóköz – ahelyett, hogy elválasztaná – éppen hogy összeköti a körülötte levő szavakat. A számítógép számára azonban mindkét szóköz ugyanazt a számkódot viseli. Azonban legyünk igazságosak: egyes szövegszerkesztők lehetővé teszik az úgynevezett nem törő szóközök bevitelét, de ezek használata nem jellemző az elektronikusan rögzített szövegekre, így a keresőgépekben sem használhatjuk fel. Egyetlen alternatívaként marad a nyelvi processzor, amely több szóra kiterjedő elemző program és szótár segítségével ismeri fel, hogy egyes szavak jobban összetartoznak, mint mások."

> Sveza dviju imenica od kojih prva atributivno određuje drugu i **ne sklanja se**[61], a obje zadržavaju svoj naglasak zove se polusloženica. Između sastavnih dijelova polusloženica piše se spojnica: *biser-grana, izvor-voda, kremen-kamen, lovor-vijenac, rak-rana, spomen-ploča, uzor-majka, plus-pol*.
> Polusloženice se rabe vrlo često u stilističke svrhe. Spojnica se kao pravopisni znak rabi i u druge svrhe (Babić-Finka-Moguš 1995: 77).

Problem se pojavljuje kod činjenice da bi se riječi morale prepoznavati kao cjelina jer se prvi dio riječi ne sklanja, te se time pojavljuje imenska paradigma s obzirom na drugu riječ, dok prva ostaje njezina sastavnica. Jedno od rješenja koje se nudi u HUMOR-u jest unošenje riječi koje se pišu spojnicom u sastavni dio rječničke osnovice programa i njihovo definiranje kao jedne cjeline. Manjkavost čini nedostatak referentnih izvora jer nisu sve moguće polusloženice sastavni dio Anićeva rječnika. Njihove se sastavnice, međutim, mogu pronaći u tekstovima kao zasebne riječi te ih time program prepoznaje. Jedno od rješenja koje se nameće jest vađenje svih onih riječi u korpusu u kojima se pojavljuje spojnica te njihovo uvrštavanje u rječničku osnovicu programa. Za taj je korak potrebna korpusna analiza kojom se dokazuje postojanje polusloženica. Testiranjem primjera iz gornjega navoda uočava se da HČR ne bilježi pojavu oblika *spomen-ploča*, ali navodi polusloženice *spomen-lokalitet* i *spomen-znak* čija je relativna čestoća pojavljivanja 0,0001 dok apsolutna čestoća pojavljivanja iznosi 1. To dovodi do zaključka da se navedene polusloženice ne pojavljuju često u korpusu, ali su one ipak zabilježene na 569. mjestu, što dokazuje njihovu uporabu te bi se kao takve morale uvrstiti u rječničku osnovicu analizatora.

Kao daljnji problem koji se tada pojavljuje, jest nedosljednost korištenja navedenih oblika. Primjerice, testiranjem hrvatskoga nacionalnog korpusa utvrđeno je naime da se pojedine leme pojavljuju u više oblika, kao što je prikazano u prikazu 17 na primjeru leme *tamnoplav*:

Oblik	Broj pojavljivanja oblika u HNK
tamnoplave	105
tamnoplavu	19
tamno plave	15
tamnoplavi	14
tamnoplavoj	12
tamnoplava	10

[61] Isticanje moje.

tamno plavoj	4
tamno plavu	3
tamno plava	2
tamno-plave	2
tamno-plava	1
tamno plavi	1
tamno-plavi	1

Prikaz 17. Izvadak iz analize oblika pridjeva *tamnoplav* iz HNK, 8. rujna 2008.

Kao što se može uočiti, broj pojavljivanja pojedinih oblika gotovo je isti za pisanje sa spojnicom ili bez nje (primjer: oba oblika, i *tamno-plavi* i *tamno plavi* pojavljuju se jedanput, dok se oblik bez spojnice *tamnoplavi* pojavljuje svega 14 puta, što nije značajan broj prema ostalim pojavnicama u korpusu). To bi ukratko značilo da bi se morfološkim parserom morale predvidjeti i paradigme navedenih oblika. Jedno od rješenja koja su pri tome moguća jest testiranje korpusa i naknadno unošenje oblika koji se pojavljuju najviše u cjelokupnome korpusu u rječničku osnovicu programa, što nije pokriveno ovim radom i ovim projektom.

O problematici koja se tiče odnosa morfologije i sintakse u ovome radu smatram bitnim napomenuti sljedeće činjenice:

1. Iz razloga što se riječ prepoznaje na grafičkoj razini, oblici koji se sastoje od više sastavnica odvojenih razmakom (primjerice složena glagolska vremena) u prvoj se fazi izrade morfološkoga parsera prepoznaju samo ako su sastavni dijelovi rječničke osnovice. Primjerice, povratno-posvojna zamjenica *sebe/se* definira se kao dio riječi u sastavu onih lema s kojima je ona navedena u Anićevu rječniku (primjerice kod leme *kupati se*). U HUMOR-u se zamjenica *se* načinom kodiranja priključuje paradigmi glagola *kupati*, čime se oblici poput *kupam se* ili *se kupam* povezuju s lemom *kupati se*, dok se primjerice u rečenici *Majka kupa dijete*. oblik *kupa* povezuje s glagolom *kupati*. To znači da se lema *kupati* dvaput unosi u rječničku osnovicu programa. Sličan se proces primjenjuje i kod negacije, gdje se, unatoč negaciji, glagol prepoznaje na odgovarajući način. Time se djelomice rješava problem razmaka kao granice analize.

2. Iako se pri izradi jednoga parsera ne mogu zanemariti sintaktičke kategorije, smatram bitnim naglasiti da sintaksa u slučaju razvoja morfološkoga parsera

predstavlja samo novi oblik definiranja pojedinih gramatičkih kategorija, poput primjerice roda imenica (više o tome u poglavlju 8.4.2.). Uz pomoć sintaktičkih kategorija paradigme će se bolje definirati, točnije opisati, te se time uvelike olakšava izrada računalnoga morfološkog rječnika. Primjenom morfološkoga parsera koji je obogaćen sintaktičkim „znanjima“ omogućit će se bolje prepoznavanje teksta te analogno time i točnija analiza korpusa.

3. Pri izradi morfološkoga parsera HUMOR-a potrebno je definirati promjenjive vrste riječi. Za razliku od tradicionalne gramatike, kao što će biti pojašnjeno u poglavlju 7., pod jednom vrstom riječi ovdje se smatraju sve one riječi koje imaju istovjetnu paradigmu. Primjerice, pod nazivom *pridjevi* u daljnjem će se tekstu podrazumijevati sve one riječi koje imaju pridjevsku paradigmu.

4. Tijekom izrade morfološkoga parsera HUMOR-a, a samim time i morfološkoga rječnika, došlo se do spoznaja koje su se riješile s tehničke strane, ali s jezikoslovne strane zahtijevaju dodatne pokuse i istraživanja. Ona se prije svega odnose na učinkovitost i primjenjivost nastalih rješenja u nastavi hrvatskoga kao drugoga ili stranoga jezika. Takva će se pitanja izričito naznačiti na odgovarajućemu mjestu. Uzimajući u obzir temu i opseg ovoga rada, nije se moglo ući dublje u rješavanje jezikoslovnih nedoumica, što će se ostvariti unutar nekoga drugog projekta. U ovome se radu u poglavlju o primjeni računalne morfološke analize daje pregled problema i plan izrade računalnoga morfološkog rječnika za učenje hrvatskoga kao drugog ili stranog jezika. Izrada demo-verzije rječnika nalazi se u posljednjoj fazi te se time smatra opravdanim dublje pozabaviti problematikom i važnošću izrade računalnoga morfološkog rječnika (više o tome u poglavlju 8.).

5. Pri morfološkoj analizi složena se slova *lj*, *nj* i *dž* smatraju jednim znakom i nisu podložna rastavljanju.

U sljedećim poglavljima slijedi detaljan prikaz načina rada morfološkoga parsera HUMOR-a te postojećih verzija morfološkoga analizatora, među njima i nepublicirane hrvatske inačice.

5. Prikaz postojećih verzija morfoloških analizatora za aglutinativne, fleksijske i visokofleksijske jezike

Prije rasprave o problemima koji se pojavljuju pri prilagodbi HUMOR-a sustavu hrvatskoga standardnog jezika, smatram bitnim dati pregled dosadašnjih verzija morfoloških analizatora s posebnim naglaskom na jezikoslovno gledište morfološke analize te odabranu problematiku. Iz razloga što se koncipiranje morfološkoga rječnika za učenje i poučavanje hrvatskoga kao stranoga jezika zasniva na HUMOR-u, većina će se primjera vezanih uz problematiku morfološke analize temeljiti na primjeru ovoga analizatora. Isto tako smatram bitnim u ovoj fazi rada predstaviti neka rješenja kojima se koriste i drugi programi, poput primjerice njemačkoga analizatora GERTWOL-a[62]. Zašto upravo uspoređivanje procesa rada HUMOR-a s GERTWOL-om? Naime, kao što je već prije navedeno, HUMOR se kao takav ugradio u različite sustave te se danas uspješno rabi za morfološku analizu engleskoga, njemačkoga, mađarskoga, poljskoga, ruskoga, francuskoga, latinskoga, starogrčkoga i turskoga jezika (postoje čak i pokusne inačice za španjolski i japanski jezik). Tvorci su HUMOR-a u više navrata naglašavali univerzalnost samoga sustava koja leži posebice u dvoslojnome modelu morfološke analize (Proszéky-Kis 1999a: 261), što potvrđuje i njemačka inačica GERTWOL-a. Parseri se u određenim fazama koriste modelom TWOL, odnosno Koskenniemijev (1983) *Two–Level–Morphology–Model* te su se tvorci obaju pasera već osamdesetih i devedesetih godina prošloga stoljeća bavili problematikom morfološke obrade mađarskoga i njemačkoga jezika. Međutim, unatoč činjenici da postoje računalne univerzalije te da se proces morfološke analize uspješno primjenjuje za analizu velikoga broja različitih jezika, postoje specifičnosti svake pojedine jezične skupine koje se ubrajaju u jezična rješenja koje programi moraju ponuditi. Važno je napomenuti da se upravo zbog specifičnosti njemačkoga računalnog modela za program GERTWOL navodi da se rabi za morfološku analizu prvenstveno njemačkoga jezika. Da bi se dakle rasvijetlila specifična problematika koja se pojavljuje kod hrvatske inačice, potrebno je promotriti određene probleme s jezikoslovnoga gledišta, odnosno jezične univerzalije koje su se pojavile pri prilagodbi analizatora različitim sustavima – sustavu aglutinativnih jezika, mađarskomu, fleksijskih

[62] http://www2.lingsoft.fi/cgi–bin/gertwol. Osim GERTWOL-a postoji i nekoliko drugih analizatora koji su se razvili za procese morfološke analize i parsinga, među najpoznatijima su programi u sustavu XEROX (XFST, TWOLC, LEXC) te MORPHY (Aleksa 2006: 143).

jezika, njemačkomu te visokofleksijskih jezika – poljskomu[63]. Osim toga, jednom od tema budućih projekata nameće se i izrada prevoditeljskih sustava za prevođenje s hrvatskoga na njemački i mađarski jezik i obratno (poput sustava MobiMouse), te stoga smatram bitnim u ovoj fazi rada suprotstaviti ova dva programa i obrazložiti odabiranje načela rada parsera HUMOR-a za računalni morfološki model analize hrvatskoga jezika.

Razvijanje morfološkoga analizatora za mađarski jezik započelo je 1992. godine, dok su neke predradnje obavljene već 1990. godine. O njemačkoj se verziji morfološkoga analizatora pisalo čak i 1982. godine (Steinacker–Trost 1982). Međutim, kada se tada razmišljalo o parsingu mađarskoga jezika s ciljem sučelja prirodnih jezika (Farkas–Naszódi 1990), te o parsingu njemačkoga jezika, razmišljalo se u oba slučaja prvenstveno o specifičnoj jezičnoj problematici koja prati razvijanje parsera, a koja će biti predstavljena u ovome poglavlju.

O specifičnostima mađarskoga jezika autori navode sljedeće:

> Mađarski je jezik takozvani aglutinativni jezik, što znači da se na korijen riječi u više slojeva mogu lijepiti nastavci, a i nekoliko se prefiksa može povezati s riječju. Osim toga i spajanje riječi bitan je način tvorbe riječi. Riječ se sastoji od funkcionalnih dijelova, koje nazivamo morfemima (*szótövek, képzők, ragok, jelek*). Određeni morfem s nekom funkcijom može imati više različitih oblika. U mađarskome jeziku dakle jedna riječ nije ništa drugo nego niz morfema (ali morfemi se ne mogu slobodno nadodavati riječi. Riječi se iz različitih aspekata ubrajaju u razrede (vidi kasnije vrste riječi, tipove deklinacije i konjugacije). Prefiksi s jedne strane mijenjaju značenje riječi, a s druge strane prenose riječ iz jednoga razreda u drugi. Zadatak je morfologije opisati s kojim se nizom morfema može stvoriti jedna riječ. Cilj je generativnoga opisa pokazati na koji se način može stvoriti jedan oblik. Zadaća analitičkoga modela jest kako u jednome obliku riječi pronaći njegove morfeme. Iz razloga što više nizova morfema mogu dovesti do istoga oblika, sve ih moramo pronaći[64] (Farkas–Naszódi 1990: 21).

[63] Iako osim poljske inačice postoji i ruska inačica spomenutoga HUMOR-a, problematika razvoja analizatora za slavenske jezike neće se razmatrati s ruskoga stajališta iz razloga što se hrvatska inačica ne mora suočiti s izborom pisma poput ruske inačice. Poljska je inačica HUMOR-a bitna jer je do sada jedina slavenska varijanta morfološkoga analizatora.

[64] Prijevod izvornika MA: „A magyar nyelv úgynevezett agglutináló nyelv, ami azt jelenti, hogy a szótőhöz több rétegben végződések tapadhatnak, és néhány előtag is kapcsolódhat a szóhoz. Ezenkívül a szóösszetétel is lényeges szóalkotási mód. A szó funkcionális részekből áll, melyeket morfémáknak nevezünk (szótövek, képzők, ragok, jelek). Egy adott funkciót megvalósító morfémának több különböző alakja lehet. A magyar nyelvben egy szó tehát nem más, mint egy morfémasorozat (de a morfémák nem kapcsolódhatnak tetszőlegesen egymáshoz). A szavak különböző szempontokból osztályokba sorolhatók (lásd később: szófajok, ragozási típusok). A toldalékok (előtagok) egyrészt a szavak jelentését változtatják meg, másrészt a szavakat egyik szóosztályból a másikba viszik át. A morfológia feladata leírni azt, hogy egy szó milyen morfémasorozattal állítható elő. A generatív leírás célja megmutatni azt, hogyan kell egy szóalakot előállítani. Az analitikus modell feladata: hogyan lehet egy szóalakból a benne szereplő morfémákat megtalálni. Mivel több morféma sorozat ugyanazt a szóalakot eredményezheti, ezeket mind meg kell találnunk.“

Tvorci njemačkoga parsera svoju problematiku opisuju na sljedeći način:

> Njemački je jezik bogat fleksijskim oblicima tako da morfološka sastavnica često rezultira s više od jednoga mogućeg stema za riječ inputa. Ovi stemovi obično pripadaju različitim kategorijama riječi, npr. 'meinen' se može interpretirati kao glagol ili se može vratiti na posvojnu zamjenicu 'mein'. Sintaksa određuje tip konstituenta, koji se očekuje na određenoj točki analize. Obično je dovoljno koristiti se sintaktičkom informacijom da bi se razriješile ovakve morfološke dvoznačnosti[65] (Steinacker–Trost 1982: 365).

Kao što se iz ovih primjera može uočiti, jedna od jezičnih univerzalija i problema koji se pojavljuju pri prilagodbi analizatora ovim dvama jezičnim sustavima odnosi se na analizu homografa te prepoznavanje stema kod nepravilnih oblika, što zalazi i u samu tvorbu riječi. Ako se pak podrobnije prouči i poljska inačica analizatora (Wołosz 2005), može se zaključiti da problem određivanja stema predstavlja izazov za tvorce računalnoga modela morfološke analize različitih jezika. Naime, Prószéky (2001: 991–992) u svome radu spominje problem homografa kao jedan od čestih problema koji se pojavljuju pri automatskoj morfološkoj analizi, odnosno analizi primjerice oblika riječi *szemtekkel* koji može imati dva značenja: *s vašim očima* i *s više smeća*. S tim u vezi program nudi dvije moguće interpretacije, odnosno povezuje se s dvama stemovima:

> szemetekkel=szemét+PL+INS
> szemetekkel=szem+PERS–PL–2+INS

Za razliku od aglutinativnih jezika, dvojbeni oblici u njemačkome jeziku nastaju zbog konverzije ili derivacije, kao što je vidljivo u sljedećem primjeru:

> Ministern → (der) Mini|stern ili (der) Minister
> Flugzeuge → (der) Flug#zeug~e ili (das) Flug|zeug
> Verbrechen → (der) Verb#rechen ili (das) Ver|brech~en (Volk 1999a: 305).

Prószéky za određivanje točne inačice koja se nudi u slučajevima dvoznačnosti, odnosno određivanja prikladnoga stema, predlaže analizu cijeloga konteksta, za što se rabe predradnje koje ne zalaze više u sastav samoga analizatora (Prószéky 2001: 991–992). Rješenja koja su primijenjena u HUMOR-u, prema riječima autora, mogu se primijeniti na

[65] Prijevod izvornika MA: „The German language is rich in inflectional forms, therefore the morphological component often comes up with more than one possible stem for an input word. These stems usually belong to different categories of words, e.g. 'meinen' can be interpreted as a verb (to suppose) or it can be reduced to the possessive pronoun 'mein' (my). Syntax restricts the type of a constituent, which is expected at a given point in the analysis. Usually it is sufficient to use syntactic information to disambiguate morphological ambiguities of this kind“

sve jezike, ali prije svega na aglutinativne. Tvorci su parsera razvili kategorije i određenu klasifikaciju koju mogu koristiti „tradicionalni lingvisti za stvaranje preciznijih definicija (ili radije za konvencije imenovanja) klasa morfema“ (Prószéky–Kis 1999a: 263).
Nijemci se u svome modelu oslanjaju na tradicionalne gramatičke kategorije i predlažu disambiguator koji se temelji na morfološkim oznakama za cijelu riječ, slabe poveznice, interfikse i derivacijske sufikse[66]. Načelom vrjednovanja svake od pojedinih oznaka i izračunavanja ukupne „vrijednosti“ pojedinoga oblika te isključivanja rijetkih oblika, Volk (1999a) je izradio sustav koji s 90%-tnom sigurnošću kod višeznačnih slučajeva oblik povezuje s odgovarajućom lemom.

Osim identifikacije stema kod homografa se velikim problemom pokazalo i određivanje stema kod nepravilnih oblika riječi. U slučaju mađarskoga jezika radi se o češćim oblicima nego u drugim jezicima, poput primjerice oblicima glagola *jőn* 'doći'-rječnički oblik (*jőnni* 'doći'-infinitiv*, jöhet* 'može doći'*, jövő* 'dolazeći'*, gyerünk* 'hajdemo') koji se mogu rastaviti na sljedeće morfeme (Prószéky 2001: 992):

jön + ni, jö + het, jöv + ő, gyer + ünk.

Samo rješenje koje predlaže HUMOR u ovome slučaju nije dostatno jer primjerice oblik *gyerünk* ne sadrži dio izvornoga stema te se na drugi način mora s njim povezati. U slučaju hrvatskoga jezika isto tako postoje oblici koji se ne mogu izravno povezati sa svojim korijenom (primjerice komparativ *bolji* s lemom *dobar*). Rješenje koje se primijenilo u slučaju hrvatskoga jezika jest da su se ti oblici posebno unijeli u rječničku bazu programa, posebnim su se oznakama na podrazini povezale s lemom te ih na taj način program može prepoznati i povezati. HUMOR se naime za rješavanje ovakvih slučajeva koristi dvoslojnim modelom (Koskenniemi 1983), zatim njegov nastavak koji je razvio Karttunen (Karttunen 1993)[67], kao i naposljetku RELEX sustav (Silberztein 1994)[68] te formalizam nastao unutar samoga parsera (Prószéky 2001: 1992).

Nadalje, ako se promotri načelo rada dvaju programa, uočavaju se sličnosti i razlike. Prema riječima autora (Haapalainen–Majorin 1994) GERTWOL nudi dva načina rada –

[66] Morfološke su oznake vidljive i u prethodno navedenom primjeru, a bit će pojašnjenje dalje u tekstu.
[67] Referenca iz (Prószéky 2001:192)
[68] Referenca iz (Prószéky 2001:192)

generiranje i analizu. Pri generiranju se kao input daje korijen riječi s poželjnim morfološkim podatcima na temelju čega program prikazuje zadani oblik riječi:

lernen+v+ind+präs+sg3 —> lernt
gut+a+komp+sg+akk+neutr —> besseres[69]

Promatrajući pak postupak analize dvaju parsera, može se uočiti da oba programa naposljetku nude isto rješenje do kojega su došli dvama različitim postupcima. Naime, oba analizatora u krajnjem rezultatu daju tradicionalne morfološke podatke o primjerice broju, rodu, padežu, načinu i glagolskome vremenu. GERTWOL radi na načelu baze podataka u obliku usmjerenoga grafa u kojemu se slaganjem znakova prema označenome inputu dobije output koji on na kraju iščita. U svome sustavu on se koristi prije svega tradicionalnim morfološkim kategorijama te ih od korijena riječi odvaja znakovima |, #, ~ i \ (Volk 1999a: 305), čime se u isto vrijeme određuju vrste veza među morfemima, a što se koristi kod disambiguatora kako je prethodno opisano. HUMOR se pak zasniva na načelu baze podataka s površinskim opisom stemova i termova i njihovim potpovršinskim kodovima koje analizira i slaže te na svakome koraku provjerava točnost navoda (Novák 2003). Razlog odabiranja dvaju pristupa morfološkoj analizi jezika iz različitih jezičnih skupina dokazuje potrebu uzimanja u obzir specifičnosti pojedinih jezika pri odabiranju računalnih morfoloških opisa. U njemačkome se slučaju zbog morfoloških zakonitosti ovakvo rješenje pokazalo učinkovitim i funkcionalnim. Kada se ovo rješenje pokušalo primijeniti na hrvatski jezik gdje kompozicija ne čini jedan od najčešćih tvorbenih modela riječi kao u njemačkome[70] te je za razliku od njemačkoga hrvatski jezik bogatiji fleksijskim morfemima, naišlo se na niz otežavajućih okolnosti pa se pokušalo pronaći neko drugo rješenje.

Iz razloga što se problemi koji se pojavljuju pri morfološkoj obradi hrvatskoga jezika (više o tome u poglavlju 6.) mogu analogno povezati s jezičnom problematikom koja je uspješno riješena pri morfološkoj obradi poljskoga jezika uslijed rješenja koja predlaže HUMOR (Wołosz 2005), odlučilo se na odabiranje načela rada ovoga parsera. Pregled specifične problematike hrvatskoga jezika koja se pojavila pri njegovoj izradi, dan je u sljedećem poglavlju.

[69] http://www.ifi.uzh.ch/CL/volk/LexMorphVorl/Lexikon04.Gertwol.html
[70] Prema Lohde (2006:63) udio je imenica u njemačkome rječniku približno 50-60% te su dva najveća tvorbena modela kompozicije i derivacije. Kompozicija se ipak pokazala u neki slučajevima relevantnija (Lohde 2006:63).

6. Izrada hrvatske inačice morfološkoga analizatora

6.1. Problematika jezične prošlosti i jezične politike pri postupku morfološke analize

Za razliku od problematike drugih jezika koji su prilagođeni sustavu automatske morfološke analize preko HUMOR-a, pri izradi hrvatske verzije veliku je ulogu imala jezična politika. Iz razloga što do sada jedina slavenska inačica ovoga analizatora s latiničnim pismom, poljska jezična inačica HUMOR-a, omogućuje između ostaloga i prepoznavanje te morfološku analizu tekstova i iz 18. stoljeća (Wołosz 2005), s pravom se postavilo pitanje treba li se to omogućiti i hrvatskoj inačici. Pri rješavanju ove dvojbe pojavilo se nekoliko problema koji se prije svega odnose na hrvatsku jezičnu prošlost i jezičnu politiku.

Prvi problem koji se pojavio jest ortografske prirode, odnosno problem jezične kodifikacije. Za razliku od drugih slavenskih jezika, kao što je već poznato, proces kodifikacije hrvatskoga jezika započeo je vrlo rano, ali je dugo trajao (Moguš 1995). Od 16. do 19. stoljeća pojavljuju se tekstovi koji su napisani nejedinstvenim grafijskim sustavom, tako da na različitim područjima Hrvatske pronalazimo više različitih varijanti za bilježenje pojedinih fonema i kodifikaciju različitih dijalektalnih inačica hrvatskoga jezika. Ovisno o području djelovanja, pojedini su autori primjerice za pisanje palatala, ovisno o utjecajima, primjenjivali talijansku ili mađarsku ortografiju. Tek se kasnije pojavljuju pokušaji ustavljivanja i korištenja ortografskoga pravila jedan glas – jedan grafem (Moguš 1995: 59).

Ako se pobliže promotre tekstovi napisani do 19. stoljeća, mogu se pronaći sljedeći grafemi koji su se koristili umjesto sadašnjih grafema[71]:

Grafemi u suvremenome hrvatskom jeziku	Neki grafemi koji su se koristili do 19. stoljeća
č	*ç, ʒ, cs, ts, cz*
ć	*c', ch, tj*
đ	*gh, dy*
lj	*l̓, ļ, ly, gl*
nj	*ñ, ń,, nj, ny, gn*
š	*ſ, sh,*
ž	*ʒ, sh, x*
–je /– ije	*Ě*

Prikaz 18: Tablica grafema koji su se koristili u 19. stoljeću

[71] Izvor za sastavljanje tablice: Moguš 1995.

Kao što se može uočiti, najvećim bi se problemom pri morfološkome opisu hrvatskoga jezika do 19. stoljeća pojavila ne samo ortografija, nego i nedosljednost njezine uporabe. Iz razloga što su se drugačija grafološka pravila primjenjivala ovisno o jezičnim inačicama i narječjima, za grafičko predočavanje jednoga fonema korišteno je nekoliko grafema. Iako postoji nekoliko rječnika iz toga razdoblja, nekoliko njih obuhvaća u svome opusu i različita hrvatska narječja, primjerice Sušnik-Jambrešićev (1742) *Lexicon latinum interpretationae illyrica, germanica et hungrica locuples*. Ipak najveći problem pri prilagodbi morfološkoga analizatora parsingu ranijih hrvatskih tekstova leži u mnogobrojnosti korištenih dijalektalnih inačica, a koje su izravno povezane i s različitim načinima pisanja, uključujući i književnu *koine* koja se razvila u 16. stoljeću (Moguš 1995: 59). Ako se nadalje promotri stanje hrvatskoga jezika s morfološke strane u vremenu do 19. stoljeća, uočavaju se različiti leksemi koji su prisutni u paradigmama kod različitih dijalektalnih inačica (Moguš 1995: 59). Devetnaesto stoljeće, nakon reforme Iliraca, donosi tekstove pisane većinom na standardiziranoj inačici, s prilično ususatvljenim pravopisom, ali s leksemima koje nisu prihvatile sve hrvatske regije, poput primjerice različitih padežnih nastavaka, kao što je nastavak *–ah* za imenički genitiv množine koji nije bio nedvojbeno prihvaćen (Moguš 1995: 170).

Sa stajališta izrade programa za morfološku analizu hrvatskoga jezika, ova saznanja donose višestruke dvojbe. Naime, kao što je već prije bilo spomenuto, program za računalnu morfološku analizu jezika morao bi prepoznavati tekstove napisane na hrvatskome jeziku i pismu. Babić (1995: 22) govori da „kao početak današnjeg hrvatskog književnog jezika može se s pravom uzeti kraj 15. stoljeća kad je štokavsko narječje u Dubrovniku ušlo u hrvatsku književnost“. Ako bismo se željeli povesti za činjenicom da morfološki analizator prepoznaje i tekstove i iz primjerice 18. stoljeća, broj bi se stemova u leksičkoj bazi programa morao uvelike povećati, te bi svaka lema primjerice s palatalima morala sadržavati i verziju s onim palatalima koji su se koristili do 19. stoljeća, odnosno svoje ranije ortografske verzije. Osim praktičnosti takve verzije programa, postavlja se i pitanje korisnosti, odnosno njegove šire primjenjivosti. Korisnike programa s velikom vjerojatnošću ne bi u tolikoj mjeri zanimalo računalno prepoznavanje tekstova iz prošlih stoljeća, a između ostaloga postavlja se i pitanje dostupnosti starijih tekstova u računalnome obliku koji bi se koristili u svrhu automatske morfološke analize. S obzirom na činjenicu da se jednim proizvodom postupka automatske morfološke analize smatra i izrada morfološkoga rječnika hrvatskoga standardnog jezika za potrebe učenja i

poučavanja hrvatskoga kao drugoga ili stranoga jezika, zamisao je o prilagođavanju morfološkoga parsera za analizu ranijih tekstova odbačena. Analizator se dakle zamislio tako da analizira samo tekstove napisane suvremenom hrvatskom ortografijom, što je za sobom povuklo neke nove dvojbe sa stajališta jezične politike i problematike leksema koji su pripadali srpsko-hrvatskome jeziku, a danas imaju svoje hrvatske istovrijednice.

S obzirom na jezični i ortografski razvoj hrvatski se jezik može podijeliti na više razdoblja. Lončarić je 1977. godine spominjao podjelu na četiri vremenska razdoblja: razdoblje od ilirizma do danas, od realizma do moderne, odnosno od početka stoljeća, koje se razdoblje poklapa i s kodifikacijom književnog jezika (Maretić, Broz-Ivekovićev rječnik), zatim razdoblje od prvoga svjetskog rata i, konačno, razdoblje od drugoga svjetskog rata do danas (Lončarić 1977: 44). Sa stajališta izrade računalnoga programa za automatsku morfološku analizu hrvatskoga standardnog jezika, s obzirom na spomenute činjenice, podjela bi se mogla ograničiti na tri razdoblja: jezično razdoblje do 1830-ih godina s neujednačenom grafijom, razdoblje do 1990-ih godina koje karakterizira korištenje srpsko-hrvatskoga / hrvatsko-srpskoga jezika, te razdoblje suvremenoga hrvatskoga jezika koje započinje 1990-ih godina. Ako se izuzme razdoblje pokušaja standardizacije hrvatskoga jezika i ideje o pobratimstvu koju je pratilo i stvaranje normirane jezične inačice, možemo reći da je stvaranje suvremenoga hrvatskoga standardnog jezika započelo 1971. godine nastavljeno 1990-ih godina, procesom jezične politike i odvajanja hrvatskoga od srpskoga jezika (Moguš 1995). Babić (1995: 17) tako navodi: „U povijesti nema zajedničkih tekstova koji su bili i hrvatski i srpski (Ima srpskih lingvista i povjesničara koji svojataju neke hrvatske tekstove, pa i cijelu dubrovačku književnost, ali to je nasilno presezanje, jednako kao što je i svojatanje teritorijalnih dijelova Republike Hrvatske)“.
Uzimajući u obzir prethodno navedene činjenice, izrada morfološkoga programa za analizu tekstova do sredine 19. stoljeća bio bi iznimno kompleksan pothvat. Prema tome, preostaje samo mogućnost prilagodbe programa ili samo suvremenom hrvatskome jeziku, ili i inačici hrvatsko-srpskoga jezika. S obzirom na korpus koji je potrebno prikupiti, vidljivo je da većina hrvatskih književnih djela iz prve polovice dvadesetoga stoljeća napisana hrvatsko-srpskim jezikom sadrži leme koje nisu više u sastavu *Rječnika hrvatskoga jezika* (primjerice lema *naredno,* koja je prema Brodnjakovu (1991: 316) rječniku ocijenjena kao srpska, u HNK se pojavljuje 65 puta, dok je nema u sastavu Anićeva rječnika). Prema riječima Stjepana Babića (1995: 23):

> Potkraj 19. stoljeća hrvatski su vukovci hrvatski književni jezik približili srpskomu, ali ni taj zahvat nije značio nestajanje hrvatske književnojezične sposobnosti. Zbog toga što norma nije stabilna, neki su pisci upotrebljavali srbizme, ali je tradicija bila tako jaka da nije mnogo prodrlo, osim kod pojedinih pisaca koji su živjeli u posebnim prilikama, kao A. G. Matoš. On je neko vrijeme živio u Beogradu pa u njegovim djelima ima mnogo srbizama, ali kao vrstan stilist nije ih sve upotrebljavao stilski neutralno. To pitanje nije proučeno pa je teško razgraničiti koliko je koje upotrebljavano s kojim stilskim vrijednostima.

Ako bi se morfološkim analizatorom trebalo analizirati ta navedena djela, bilo bi potrebno proširiti postojeći opis u smislu unošenja novih riječi i stemova te uključiti riječi koje danas pripadaju izričito srpskom jeziku, što otvara nove jezikoslovne dvojbe.

Jezik do 1990-ih godina, srpsko-hrvatski, razlikuje se od današnjega hrvatskoga standardnog jezika, kako na morfološkome, tako i na sintaktičkome i leksičkome polju. I sa stajališta jezične politike hrvatski se i srpski jezik danas smatraju jezicima, a ne dijalektalnim inačicama jednoga jezika[72]. Kritički se prema hrvatskoj jezičnoj politici odnosi i Babić koji govori: „Danas kada se hrvatski književni jezik slobodno normira, postavlja se pitanje njegova odnosa prema srpskome književnom jeziku. Taj se problem ne može mimoići već zato što danas kod mnogih prevladava mišljenje da se treba što više udaljiti od srpskoga, da je to nacionalni zadatak i mnogi tako postupaju ili žele da se tako postupa“ (Babić 1995: 29). Međutim, unatoč tomu, još uvijek pojedini jezikoslovci u nekim izdanjima tvrde da hrvatski i srpski jezici nisu zasebni jezici, nego samo dijalekti istoga, srpsko-hrvatskoga jezika, poput Wardhaugha (1991). U svome djelu Wardhaugh (1991: 29) tvrdi da glavna razlika između hrvatskoga i srpskoga jezika leži u različitim leksemima kojima se oni koriste i da ne postoji razlika na jezičnoj i fonetskoj razini između tih dvaju jezika. Za razliku od Wardhaugha, hrvatski su jezikoslovci mišljenja da je razlika između tih dvaju jezika očita na svim područjima i uključuje razlike na morfološkoj, sintaktičkoj, semantičkoj i fonetskoj razini (Težak 2004). Ovu činjenicu dokazuju između ostalog mnoga izdanja napisana nakon 1991. godine, a koja su pridonijela kodifikaciji suvremenoga hrvatskog jezika i koja ističu razlike između hrvatskoga i srpskoga jezika, poput primjerice Brodnjakova (1991) *Rječnika razlika*

[72] Iz razloga što tema ovoga rada nije hrvatska jezična politika i problemi normiranja hrvatskoga i srpskoga jezika, ovom se problematikom neće detaljnije baviti, nego se preuzimaju promišljanja hrvatskih jezikoslovaca. Teze se temelje na odabranoj literaturi – izdanjima Babića (1995) i (2004) i Težaka (1991), (1995), (1999a), (2004).

između hrvatskoga i srpskoga jezika[73] (detalji o rječniku dani su u poglavlju 2.2.). Ova i slična izdanja mogla bi se smatrati referentnim izvorom pri razgraničavanju ovih dvaju jezika te pri izradi morfološkoga analizatora za parsing hrvatskoga standardnog jezika, da se ne pojavljuju određeni problemi o kojima će biti više riječi u sljedećem odlomku.

Jedan od najvećih problema koji se pojavljuje pri testiranju programa na hrvatskome jezičnom korpusu sastavljenom za tu prigodu (više o korpusu u poglavlju 6.3.) jest činjenica da su pojedini leksemi koji su prije pripadali rječničkome opusu srpsko-hrvatskoga jezika (a danas se smatraju izričito dijelom srpskoga jezika sa svojim hrvatskim istovrijednicama) još uvijek prisutni u svakodnevnome razgovornom i pisanome jeziku. Primjerice lema *stepenice*, koja se prema Brodnjakovu rječniku smatra danas dijelom srpskoga jezika sa svojom istovrijednicom *stube* u hrvatskome jeziku, još je uvijek prisutna u Anićevu rječniku te u pisanim medijima (na hrvatskim internetskim stranicama prisutna je u 92 100 slučajeva[74], a osvanula je i 2008. godine na reklamnome plakatu u izrazu „Pad niz stepenice ne ostavlja trag vjenčanog prstena“[75]). Međutim, Brodnjakov se rječnik ne smije kategorički shvatiti. Babić, pojašnjavajući njegovu uporabu, navodi:

> Zbog boljega razumijevanja moram ponoviti ono što sam već i rekao i Brozović napisao: ne valja misliti da je lijeva strana samo srpska, a desna hrvatska jer je i lijeva strana jednim dijelom hrvatska. Zato lijevu stranu valja shvatiti ne kao srpsku, nego kao nehrvatsku, manje hrvatsku, slabije hrvatsku ili dobru hrvatsku gdje je na desnoj strani još bolji hrvatski izraz i bar jednakovrijedan koji Srbi nemaju (Babić 1995: 53).

Ako slijedimo ovaj navod, postavlja se problematika ograničavanja ovih dvaju jezika, odnosno pitanje vjerodostojnosti izvora.

Uzimajući u obzir jezičnu uporabu prema navedenom primjeru, postavlja se pitanje referentnih izvora, jezične politike odnosno sastavljanja rječničke baze morfološkoga analizatora, a time i samoga morfološkog rječnika. Naime, dvojba se prije svega odnosi na svrhovitost morfološke analize, a time i morfološkoga rječnika koji bi trebao u što boljoj mjeri predstavljati leksik koji se rabi na hrvatskome govornom području i koji je dio hrvatskoga jezičnog korpusa te učenicima davati pomoć pri svakodnevnoj komunikaciji i na hrvatskome jezičnom području. Prema mišljenju Jure Šonje „Broz-Ivekovićev je

[73] Dalje u tekstu: Razlikovni rječnik
[74] Rezultati pretraživanja google internetskim pretraživačem 27. lipnja 2008. godine.
[75] Reklamni je plakat uočen u Osijeku 16. listopada 2007. godine.

Rječnik bio srpskohrvatski i folklorni premda je u naslovu nosio hrvatsko ime. Taj manjak, prema Šonji, nije ispravio ni Anićev hrvatski jednojezični rječnik, koji je doduše moderniji, ali i on obiluje srpskim riječima“ (Danolić 2000: 44). Ako se uzme ovaj podatak u obzir, s pravom se postavlja pitanje vjerodostojnosti izvora koji stoje na raspolaganju pri izradi morfološkoga analizatora za hrvatski standardni jezik. Jer ako se na hrvatskome jezičnom području unatoč naporima hrvatskih jezikoslovaca još uvijek koriste leksemi koji se danas smatraju dijelovima leksika srpskoga, odnosno bivšega srpsko-hrvatskoga jezika, nameće se problematika koncipiranja i leksičke baze morfološkoga rječnika.

Ako bi HUMOR sadržavao samo morfološki opis sastavnica hrvatskoga standardnog jezika, odnosno ako bi u leksikon ulazile samo leme iz Anićeva (2000) ili Šonjina rječnika, koje ne pripadaju leksiku bivšeg srpsko-hrvatskoga jezika, odnosno ako bismo u obzir uzeli i Brodnjakov *Razlikovni rječnik*, program ne bi prepoznavao riječi u korpusu koje su nekada bile dijelom hrvatsko-srpskoga leksikona, a danas pripadaju srpskom jeziku, pri čemu se smanjuje relevantnost sustava i njegova pouzdanost. Točnije rečeno, tekstovi napisani do 1990.-ih godina ne bi bili u potpunosti prepoznati, a time ni morfološki analizirani. To se odnosi prvenstveno i na djela koja se danas ubrajaju u hrvatsku književnost, a napisana su prije 1990.-ih godina te sadržavaju poneke leme koje danas nisu u sastavu hrvatskoga rječničkog blaga. Dvojba koja se sada ovdje postavlja odnosi se prvenstveno na sastavljanje „purističke“ inačice morfološkog analizatora, koji bi bio u mogućnosti morfološki analizirati samo tekstove napisane suvremenim hrvatskim jezikom, odnosno tekstove nastale nakon osamostaljivanja Hrvatske i kodificiranja suvremenoga hrvatskoga jezika, ili bi ipak trebalo proširiti leksičku bazu HUMOR-a lemama koje se pojavljuju u hrvatskim tekstovima i djelima, a pripadale su bivšem srpsko-hrvatskom jeziku te se smatraju dijelom srpskoga leksika. Ako se pak leksička baza HUMOR-a proširi takvim sastavnicama, postavlja se pitanje može li se onda ova inačica programa nazvati hrvatskom. Pomoć pri rješavanju ove dvojbe može se potražiti u rješenju koje je primijenio Moguš pri sastavljanju svoga čestotnika. Vidimo da je i ondje bila prisutna problematika izbora jezične inačice te stoga autori (Moguš-Bratanić-Tadić 1999a: 6) navode:

> Omeđili smo dakle korpus na tekstove u rasponu od 45 godina veoma plodnoga rada hrvatskih književnika. Nije bilo sumnje da to jest hrvatski književni jezik. Time je ujedno bila riješena dilema koja se mogla postaviti: naime, treba li naš rječnik obuhvatiti građu cjelokupnoga „hrvatskoga ili srpskoga jezika“ ili samo građu tzv. „hrvatske varijante“. Čak i oni koji su smatrali, doduše, najčešće

> teoretski, da bi bilo „idealno kad bi se paralelno mogli izraditi čestotnici svih standardnih idioma kojima je osnova štokavsko narječje“ jer bi to „pružalo mogućnosti zanimljivim usporedbama i saznanjima“ sugerirali su ipak da je u „našim prilikama“ bolje „izraditi čestotnik za jedan od tih kodova koji nama pripada, tj. čestotnik hrvatskog književnog jezika“ (Moguš-Bratanić-Tadić 1999a: 6).

Ako bi se ovo poimanje jezičnih varijanti uzelo u obzir, možemo zaključiti da su, unatoč činjenici da su književni i standardni jezik sinonimi kako se navodi u Anićevu rječniku[76], autori odvojili ta dva pojma te napravili korpus hrvatskoga književnog jezika. Jonke (2005) pojašnjava terminologiju te navodi:

> Da bi se uklonila mogućnost nesporazuma, u nekih se naroda književni jezik naziva kulturnim jezikom ili jezikom kulture. Neki ga lingvisti nazivaju i standardnim jezikom jer sve u njemu treba da bude određeno i jasno, uvijek naime treba sasvim sigurno znati što koji znak ili riječ pokriva po značenju. Književni je jezik svakako jezik neke narodne zajednice koji stoji iznad dijalekata i sposoban je da bude komunikativno sredstvo za pripadnike raznih dijalekata. To nije samo jezik književnosti, nego jezik čitave narodne kulture. (Jonke 2005: 26)

Korpus po svemu sudeći može sadržavati leme koje su se prije smatrale sastavnicama srpsko-hrvatskoga jezika, a danas su u čestoj uporabi unatoč činjenici da pripadaju srpskom jeziku.

Jure Šonje, autor *Rječnika hrvatskoga jezika*[77]*,* ovu je problematiku riješio na način da je u rječnik uvrstio riječi kojima je izvor srpski, jer je to po njegovu mišljenju posljedica dugogodišnje jezično-unitarističke teorije te života u zajedničkoj državi. Sukladno tomu on govori:

> Na upit jesu li u Rječnik uvrštene i srpske riječi njegov autor odgovora potkrepljujući svako objašnjenje nizom podataka. Kaže da je izbor natuknica rađen po statusu riječi u hrvatskome jeziku. Srpskih riječi uglavnom nema, jer one ne pripadaju hrvatskomu jeziku. Rječnik se pridržava načela da hrvatski i srpski jezik nisu jedan jezik, nego dva jezika. Doduše, od svih slavenskih jezika hrvatski jezik najbliži je srpskomu i slovenskomu jeziku: srpski hrvatskim štokavcima, a slovenski hrvatskim kajkavcima, drži Šonje. U hrvatskome jeziku postoji znatan broj riječi čiji su izvori srpski. To je posljedica dugogodišnje jezično-unitarističke teorije i prakse hrvatskih jezikoslovaca od kraja 19. st. do najnovijih dana i života u zajedničkoj državi.

[76] Izvadak iz Anićeva (2000) *Rječnika hrvatskoga jezika*: „standardni jezik – *lingv.* normirani i kodificirani opći jezik opće pismenosti (jezik školstva, uprave, kulture, sredstva informiranja i dr.); književni jezik.“; „književni jezik – 1. standardni jezik“

[77] Šonje, J. (2000.) Rječnik hrvatskoga jezika, Zagreb: Grafički zavod Hrvatske

> – Stoga sam sve takve riječi verificirao prema Akademijinu korpusu starih hrvatskih pisaca i leksikografa. Riječi koje nemaju potvrdu u tome korpusu uputio sam na one koje imaju, a ako takvih nema, onda sam ih obradio kao dio hrvatskoga korpusa, jer to traže zahtjevi jezične funkcionalnosti. Tako npr. leksikografska odrednica kao uputnica primjenjuje se kod srpskih riječi ili oblika koji se javljaju i u govornika hrvatskoga jezika, ali se ne navodi da su riječi srbizmi. Srpske riječi ili srpska značenja zajedničkih riječi uopće se ne donose ako nisu često u uporabi u hrvatskome književnom jeziku ili ako su većini govornika hrvatskoga jezika prepoznatljive kao srpske (*vazduh, prevoshodno, opšte*). Šonje se služio Guberina–Krstićevim *Razlikama* i Brodnjakovim *Razlikovnim rječnikom*, ali i prema njima je bio kritičan (…)
>
> Uklapaju li se *sistem* i *prisutnost* u Vaše jezične poglede, pitamo gospodina Šonju. Njegov Rječnik preferira *sustav* i ne progoni *sistem*, koji je u proteklom desetljeću bio nepodoban i gotovo je nestao iz hrvatskoga jezika. Uz natuknicu *sistem*, bez definicije i primjera, Rječnik donosi sinonim *sustav* s definicijom i primjerima. Slično je i s riječju *prisutnost*. Po njegovu sudu hrvatskomu jeziku potrebne su obje riječi, *prisutnost* i *nazočnost* (Danolić 2000: 44).

Kada bi se prihvatilo rješenje primijenjeno u rječnicima hrvatskoga jezika i kada bi se sve leme iz rječnika uključile u leksičku bazu parsera, omogućilo bi se po svemu sudeći prepoznavanje tekstova napisanih do devedesetih godina prošloga stoljeća, što dovodi do sljedeće dvojbe koja se odnosi na sastavljanje morfološkoga rječnika koji se temelji na morfološkome analizatoru.

Ako bi u sastav morfološkoga analizatora ušle one leme koje se danas ne smatraju dijelom hrvatskoga leksikona, i ako bi se rječnička baza automatski preuzela za sastavljanje morfološkoga rječnika, onda se sastavljeni rječnik ne bi mogao nazvati hrvatskim. S druge strane, ako su se u pojedinim sferama jezika (pisanome, razgovornome ili jeziku mladih) još uvijek u velikoj mjeri zadržale nehrvatske leme koje se smatraju ostavštinom srpsko-hrvatskoga jezika, i ako je postavljeni cilj izrade morfološkoga rječnika osim poučavanja i učenja hrvatskoga kao drugoga ili stranoga jezika i olakšavanje komunikacije na hrvatskome jeziku, postavlja se pitanje u kojoj bi mjeri čisto hrvatski morfološki rječnik ispunio postavljene ciljeve. Isto se tako postavlja pitanje izabiranja hrvatskoga književnog jezika kao jedne od mogućih inačica te njezina dostatnost pri sastavljanju morfološkoga rječnika hrvatskoga jezika. Odgovor na ovo pitanje može dati korpusna analiza testnoga korpusa.

Kao zaključak problematike jezične prošlosti i jezične politike o kojoj se elaboriralo u ovome dijelu rada, donosi se rješenje o sastavljanju rječničke baze analizatora na temelju Anićeva rječnika s mogućnošću njenoga proširivanja lemama koje se često pojavljuju u

tekstovima testnoga korpusa, a koje nisu dijelom hrvatskoga rječnika, odnosno onim lemama koje nisu u sastavu Anićeva rječnika. Prilikom sastavljanja rječničke baze morfološkoga analizatora kao referentni izvor uzet će se Brodnjakov *Razlikovni rječnik* te će se umjesto dvojbenih lema koristiti hrvatski ekvivalenti, kao i prilikom davanja prijevoda natuknica iz stranih jezika.

6.2. Problematika jezične inačice pri izradi morfološkoga analizatora

Slično kao što je opisano u poglavlju 3.2., koje se odnosi na problematiku jezične inačice pri izradi morfološkoga rječnika, može se promatrati i problematika izbora jezične inačice pri izradi morfološkoga analizatora.
Kao što je već objašnjeno, morfološki se analizator izradio za morfološku analizu pisanih tekstova, odnosno implementiranje u različite sustave za učenje i poučavanje stranih jezika. Iz razloga što je jedan od tih sustava i sustav za prevođenje, postavlja se pitanje izbora jezične inačice.

Ako bi se HUMOR primijenio samo u sustavu prevođenja pisanih tekstova, odnosno riječi koje se pojavljuju na korisnikovu zaslonu, problematika se izbora jezične inačice sužava, odnosno postaje logičnim prilagodba sustava analizi samo pisanih tekstova na standardnoj inačici hrvatskoga jezika. Međutim, iz razloga što bi se trebalo omogućiti i daljnje razvijanje HUMOR-a, odnosno mogućnost prilagodbe sustava prepoznavanju i razgovornoga jezika, postavlja se pitanje uključivanja prozodijskih karakteristika u leksičku bazu cjelokupnoga sustava te naposljetku dijalektalnih inačica. Jedan od problema koji se pri tome pojavljuju jest nedostatak referentnoga materijala za takav pothvat, odnosno nedostatak materijala koji bi ušao u testni korpus za testiranje vjerodostojnosti programa.

Ako se pobliže promotre leme u Anićevu rječniku, možemo vidjeti da rječnik sadrži oznake za naglasak kod pojedinih riječi. Problem se međutim pojavljuje kod referentnih pisanih izvora za pisanje naglasaka kod različnica jedne natuknice. Primjerice, iako kod leme *noć* postoje oznake za naglaske čak i u dvojbenim slučajevima (u genitivu jednine, instrumentalu i genitivu množine), nedostaje pouzdani izvor za bilježenje naglasaka na

ostalim leksemima u paradigmi, odnosno referentni izvor za njihovo navođenje. Ako se nadalje pobliže pogleda *Gramatički tezaurus,* može se uočiti da ni on ne daje prozodijske elemente, nego samo navodi paradigmu te se usredotočuje na morfološke karakteristike riječi. Iz ovih se navedenih razloga odbacio način uvođenja prozodijskih elemenata u leksičku bazu programa, što podržava i činjenica da se po potrebi leksička baza HUMOR-a može pretvoriti u oblik koji će služiti prepoznavanju govora. Parser se naime već uspješno koristi kod prepoznavanja govornoga mađarskoga jezika u okviru sustava *AMor* (Acoustics & Morphology), o čemu autori navode sljedeće:

> Pretpostavlja se da se atomski segmenti fonetskoga signala za input sastoje od (podspecifiranih) fonema ili fonemskih skupina. Da bi se dobila baza podataka s fonetskim opisom, već postojeća (ortografska) baza podataka mora se konvertirati. Ova konverzija nije tako jednostavna kako se isprva čini. Svi se ortografski oblici moraju konvertirati u što je više moguće fonetskih alomorfa na temelju konverzije tipa „grafemska sekvenca u fonetsku sekvencu". Ovakav set sadrži svaki alomorf, bilo da se radi o prvome ili posljednjemu fonemu koji je podložan kontekstualnoj promjeni[78] (Prószéky 1997: 144).

Kao što se može zaključiti, iz razloga što je konverzija i nakon sastavljanja leksičke baze analizatora moguća, te iz razloga što je cilj ovoga rada prvenstveno oblikovanje morfološkoga rječnika koji će služiti za učenje i poučavanje hrvatskoga kao drugoga ili stranoga jezika, u ovoj se fazi projekta odbacila ideja o bilježenju naglasaka na lemama u leksičkoj bazi parsera.

Druga dvojba koja se pojavila odnosila se na morfološku analizu i ostalih inačica jezika osim standardne varijante, što prvenstveno uključuje razgovorni jezik. Problem koji se pri tome pojavio odnosio se na uvođenje elemenata razgovornoga jezika u leksičku bazu analizatora, odnosno ponovno se pojavio problem referentnih izvora. Iz razloga što je testni korpus sastavljen na temelju Lončarićeve (1977) preporuke u kojoj se navodi: „Svuda su zastupljene četiri glavne stilske skupine: govorni jezik ili dijalozi reprezentirani dramskim djelima, dakle fingirani govor, 2. beletristika, 3. žurnalistika i 4. znanstvena proza" (Lončarić 1977: 44), te sadrži sastavnice govornoga jezika (dramske tekstove, koji su

[78] Prijevod MA, izvornik: „Atomic segments of the phonetic input signal are assumed to consist of (underspecified) honemes or phoneme complexes. In order to get a database of phonetic description of stems and suffixes, the already existing (orthographical) database has to be converted. The conversion is not as simple as it sounds first. All the orthographic representations need to be converted into as many phonetic allomorphs as can be made on the basis of a grapheme–sequence–to–phoneme–sequence conversion. This set contains every allomorph either the first or the last phoneme of which is subject to contextual change".

najbliži kodifikaciji razgovornoga jezika), može se zaključiti da su sastavnice razgovornoga jezika uključene u leksičku bazu analizatora. Analizator se naime testira na testnome korpusu s ciljem uvođenja sastavnica i lema koje se ne nalaze u leksičkoj bazi programa, a koja je napravljena na temelju Anićeva rječnika. Sastavnice razgovornoga jezika prisutne su ne samo u dramskim tekstovima, nego i u modernijim oblicima internetskih stranica – blogovima, koje karakteriziraju, osim sastavnica razgovornoga jezika, i elementi žargona, odnosno idiolekt mladih te neformalni odnosno razgovorni stil pisanja (više o tome u poglavlju 6.3.). Iz razloga što je HUMOR fleksibilni sustav koji omogućuje proširivanje svoje leksičke baze, na temelju testiranja odlučuje se koji će se elementi unijeti u bazu podataka da bi se povećala prepoznatljivost samih tekstova. Ako se pojedini elementi često pojavljuju u testnome korpusu, oni će se uvrstiti i u morfološki leksikon.

Kao sljedeća dvojba koja se pojavila pri izboru jezične inačice bilo je inkorporiranje i dijalekata odnosno dijalektalnih sastavnica u rječničku bazu programa. Iako je već u prethodnome poglavlju pojašnjena odluka za izabiranjem standardne jezične inačice za sastavljanje morfološkoga leksikona, ostalo je pitanje uključivanja i dijalektalnih inačica u sastav morfološkoga analizatora. Razlog ovoj dvojbi pojavio se nakon činjenice da se na internetskim stranicama (posebice blogovima) mogu pronaći i tekstovi napisani na dijalektalnim inačicama, s elementima razgovornoga jezika. Iz razloga pak što je broj tekstova koji se pojavljuju na dijalektalnim inačicama prilično mali u usporedbi sa standardnim jezikom, pri njihovu uključivanju u leksičku bazu programa povećao bi se uvelike broj stemova i termova, a općenita prepoznatljivost tekstova na hrvatskome jeziku ne bi uvelike porasla.
Iz svih navedenih razloga odlučilo se dakle na oblikovanje analizatora za računalnu morfološku analizu pisanih tekstova na hrvatskome standardnom jeziku.

6.3. Testni korpus

Kao što je već bilo spomenuto, da bi se program verificirao, bilo je potrebno sastaviti testni korpus za provjeru rada parsera. Iz razloga što se ovdje radi o morfološkoj analizi tekstova pisanih na hrvatskome standardnom jeziku, bilo je potrebno sastaviti reprezentativan

korpus na temelju čijeg će se testiranja pratiti dostignuća programa te ispraviti mogući nedostatci.

Za početak se pošlo od definicije korpusa, prema definiciji u Tadić (1998), gdje se među ostalim razlikuju sljedeće kategorije:

> – zbirka tekstova: svaki skup tekstova koji je skupljen prema nekim kriterijima
> – korpus: zbirka jezičnih odsječaka koji su odabrani i skupljeni prema eksplicitnim lingvističkim kriterijima upravo s ciljem da čine jezični uzorak2
> – računalni korpus: korpus koji je kodiran na standardan i dosljedan način s nakanom da bude otvoren za računalno pretraživanje (Tadić 1998: 337).

Zaključilo se da se za potrebe testiranja morfološkoga parsera treba započeti sa sastavljanjem jednoga korpusa, koji bi po svemu sudeći trebao biti reprezentativan. Iz razloga što se ovdje radi o testnome korpusu za potrebe morfološke analize, neće se odlučiti za računalni korpus, nego za korpus u tekstualnom formatu koji je podoban za potrebe pretraživanja. Nadalje, postavilo se pitanje reprezentativnosti jednoga korpusa, odnosno izbora samih tekstova koji će ući u njegov sastav. Prema Lemnitzer-Zinsmeister (2006) ne postoji kategorija reprezentativnosti korpusa, nego korpus može biti samo jedan uzorak neke temeljne ukupnosti koja postoji u jeziku. Iz razloga što se odnos uzorka prema temeljnoj ukupnosti u sadašnjemu jeziku ne može točno definirati[79], ne možemo ni govoriti o kategoriji reprezentantnosti nekoga korpusa, negp samo približavanju idealnom stanju[80] (Lemnizter-Zinsmeister 2006: 50). Lemnitzer i Zinsmeister nadalje određuju kategoriju reprezentatnosti korpusa sljedećim navodom:

> Kao što smo prethodno definirali, korpus je uvijek jedna vrsta pokusnoga uzorka, o kojoj ne znamo je li uistinu reprezentatna i predočava li odnose kao što su oni u cjelini. Ova činjenica međutim ne onemogućuje da se na temelju podataka u korpusu dođe do lingvističkih saznanja[81] (Lemnizter–Zinsmeister 2006: 54).

[79] Autori to pojašnjavaju na sljedeći način: „Bei einer gegenwärtig verwendeten Sprache können wir das Verhältnis von Stichprobe zu Grundgesamtheit nicht exakt bestimmen“ (Lemnizter–Zinsmeister 2006: 51).
[80] Prema riječima autora : „Dahinter steckt die Frage, inwieweit man Erkentnisse, die man durch die Analyse von Korpusdatenn gewonnen hat, auf den Sprachauschnitt, den das Korpus repräsentieren soll, übertragen kann (…) In der Terminologie der Statistik spricht man davon, dass die Grundgesamtheit, über die man etwas aussagen möchte, nicht präzise definiert werden kann“ (Lemnizter–Zinsmeister 2006: 50).
[81] Prijevod MA. Izvornik:“ Wie wir oben festgestellt haben, ist ein Korpus immer nur eine Art Stichprobe, von der wir nicht wissen, ob sie wirklich repräsentativ ist und die Verhältnisse so widerspiegelt, wie sie auch in der Gesamtheit sind. Diese Tatsache verhindert aber nicht, dass man linguistische Erkenntnisse über eine Sprache anhand von Korpusdaten gewinnt.“

Nakon definiranja reprezentatnosti korpusa prešlo se na izbor tekstova koji će se uvrstiti u korpus. Pri sastavljanju testnoga korpusa postavilo se isto tako pitanje kategoriziranja tekstova koji će se uvrstiti u korpus. Pojedina su pitanja već pojašnjena u prethodnome poglavlju, međutim, iako se odlučilo na odabiranje tekstova napisanih na standardnome jeziku, potrebne su bile neke smjernice kod izbora samih tekstova. Lemnitzer i Zinsmeister (2006) pri tome daju sljedeći naputak:

> Ravnoteža se ovdje odnosi ponajprije na vrste tekstova. Putanja prema jednom uravnoteženom korpusu postiže se zajedničkim djelovanjem unutarnjih i vanjskih kriterija. Nadalje se vrste iskaza odabiru prema eksternim kriterijima, primjerice prema broju uključenih osoba (govor, intervju, dramski tekst itd.), prema stupnju usmenosti ili pismenosti (spontani govor, pročitani govor, chat-protokol, novinski članak itd.), prema situaciji (formalni, neformalni, itd.)[82] (Lemnizter-Zinsmeister 2006: 52).

Slično govori i Tadić (1998) koji, precizirajući sastav reprezentativnoga korpusa, navodi:

> Jedno je sigurno, ako se želi korpus reprezentativan za neki jezik, pri njegovu se sastavljanju mora paziti na:
> – različita područja uporabe jezika (teme, discipline)
> – tipove tekstova (knjige, novine, časopisi, brošure, prospekti, pisma itd.)
> – dužina tekstova (knjige, pripovijetke, crtice, članci)
> – žanrove (lijepa književnost, publicistika, znanost, udžbenici, novinski tekstovi itd.)
> – medij ostvarivanja jezične poruke (pisani, govoreni jezik)
> – autorove osobine (dob, spol)
> – vrijeme nastanka teksta (u našem slučaju od 1990. do dana zaključenja korpusa) (Tadić 1998: 341).

Iz razloga što je testni korpus za testiranje morfološkoga analizatora, kao što je već definirano i utvrđeno, samo jedna vrsta pokusnoga uzorka koja bi trebala odgovarati većoj cjelini, smatralo se potrebnim testni korpus uskladiti s Hrvatskim nacionalnim korpusom[83]. Na taj se način dobivaju relevantni podatci, što omogućuje točnije sastavljanje morfološkoga analizatora, a time i morfološkoga rječnika. Sljedećim se korakom time podrazumijeva proučavanje sastava HNK-a.

[82] Prijevod MA. Izvornik: „Die Ausgewogenheit wird hier vor allem auf Textsorten bezogen. Der Weg zu einem ausgewogenen Korpus soll durch das Zusammenspiel von externen Kriterien und internen Kriterien erreicht werden. Zunächst werden Äußerungssorten nach externen Kriterien ausgewählt, z.B. nach der Anzahl der beteiligten Personen (Rede, Interview, Schauspiel etc.), nach dem Grad der Mündlichkeit und Schriftlichkeit (spontanes Gespräch, abgelesene Rede, Chatprotolokk, Zeitungsartikel etc.), nach der Situation (formell, informell, etc).“

[83] Dalje u tekstu HNK

Kao što se iz sljedećeg citata može pročitati, temeljne odrednice HNK-a u suglasnosti su s postavljenim ograničenjima HUMOR-a, što znači da su u korpus ušla djela napisana suvremenih hrvatskim jezikom:

> Ako se danas želi sastaviti korpus hrvatskoga suvremenog jezika, onda se može poći od točke u vremenu koja je u mnogome prijelomna ne samo za hrvatski jezik već i za Hrvatsku kao državu i Hrvate kao narod. Riječ je, dakako, o godini 1990. Stoga bi se korpus suvremenoga hrvatskoga jezika (KSHJ nadalje) valjalo započeti s godinom 1990. Netko bi mogao prigovoriti da je takva odluka u potpunosti nelingvistička, gotovo politička. No čini se da se jezičnih argumenata za takvu odluku može naći jer svi, dakako intuitivno, osjećamo da smo od tada hrvatski mogli rabiti „slobodnije", „spontanije" ili, gotovo pjesnički rečeno, mogli smo ga konačno „disati punim plućima" (Tadić 1998: 338).
>
> Da ne bi bilo zabune: hrvatski je postojao kud i kamo prije te godine, ali korpusi, osobito suvremenoga jezika, moraju se ograničiti i započeti od neke točke u vremenu. Sve što je na hrvatskome nastalo prije 1990. može se uključiti u korpus koji se ne bi zvao korpus suvremenoga hrvatskoga jezika već bi pripadao u Zbirku hrvatskih tekstova (Tadić 1998: 339).

Tekstovi u Hrvatskom nacionalnom korpusu podijeljeni su na pet potkorpusa – drama, novine, proza, stihovi i udžbenici. Veličina je uzoraka, prema riječima autora, bila između 5, 10 i 20 tisuća pojavnica, ali se odabrao uzorak od 10 000 riječi kontinuiranog teksta svakoga od navedenih autora (Moguš-Bratanić-Tadić 1999: 7) da bi se postigla bolja raspršenost uzoraka prema različitim autorima[84].

Precizniji izbor pisanih tekstova koji su ušli u sastav HNK-a jest sljedeći:

a) pisani tekst
Knjige
- Proza (romani (povijesni, kriminalistički...), pripovijetke, crtice, dnevnici, eseji)
- Publicistika (knjige, članci, kronike)
- Znanost (knjige, rasprave, članci raznih struka)
- Udžbenici (srednjoškolski i sveučilišni udžbenici raznih struka)
- Priručnici (tehnički, kulinarski, odgoj djece, domaćinski...)
- Zakoni (Narodne novine, zakonski tekstovi, pravnički časopisi)

Novine
- dnevnici
  - nadregionalni
  - regionalni
- tjednici
- dvotjednici

Časopisi

[84] http://www.hnk.ffzg.hr/struktura.html, 22.srpnja 2008.

– tjednici
– dvotjednici
– mjesečnici
– višemjesečnici
Brošure, prospekti
Korespondencija
– privatna
– službena
b) govoreni tekst:
formalni/pripremljeni (predavanja, izlaganja, nastupi)
neformalni/ad hoc (dijalozi npr. RTV i druge javne diskusije itd.)[85]
(Tadić 1998: 351)

Nakon definiranja sastavnica testnoga korpusa prionulo se prikupljanju tekstova koji će ući u njegov sastav. Iz razloga što se ovdje radi o testnome korpusu u opsegu od 5 milijuna riječi koji je sastavljen za potrebe istraživanja odnosno testiranja već postojećega programa, a ne o korpusu koji je nastao unutar jednoga projekta, moralo se voditi računa i o autorskim pravima. Da bi se međutim pokrila sva područja predodređena HNK-om, napravljeno je sukladno s Hrvatskim nacionalnim korpusom pet potkorpusa koji se sastoje od sljedećih kategorija:

1. najvažnija književna djela u čijemu su sastavu najvažnija djela iz hrvatske književnosti (drame, novele i pjesme),
2. efemerna građa te tiskane publikacije (novinski članci, članci iz časopisa),
3. tekstovi s elektronskih medija (internetske stranice, blogovi),
4. tekstovi iz stručne literature,
5. tekstovi iz različitih udžbenika.

Izbor tekstova koji će ući u pojedine žanrove temeljio se na odrednicama koje su pojašnjene i u prethodnome poglavlju. Dvojba koja se odnosila na izbor i rangiranje najvažnijih književnih djela riješena je korištenjem kompilacije najvažnijih hrvatskih književnih djela naklade Bulaja (1999a, 2000, 2002) *Klasici hrvatske književnosti*. Iako kompilacija sadrži djela koja su napisana od 16. do 20. stoljeća, u uži su izbor, kako je pojašnjeno, ušla samo djela napisana standardnim hrvatskim jezikom.
Drugi potkorpus koji sadrži efemernu građu i publikacije odražava suvremenu jezičnu situaciju. Novinski tekstovi iz različitih su izvora – mjesečnika, tjednika, koji pokrivaju

[85] Za točan popis tekstova i pisaca koji su ušli u sastav Hrvatskoga nacionalnog korpusa vidjeti Moguš-Bratanić-Tadić (1999: 7-10).

široki spektar registra – jezika časopisa za žene, muškarce i tinejdžere. Cilj je sastavljanja ovoga potkorpusa bio pokrivanje što većeg raspona jezičnih inačica i registara.
U treći potkorpus ušli su tekstovi s elektroničkih medija, internetskih stranica i blogova. Kada se ovi izvori uzimaju u obzir, mora se obratiti pozornost na sljedeće stvari:

1. Ne sadrže sve internetske stranice tekstove napisane hrvatskim standardnim jezikom, nego mnoge stranice sadržavaju još uvijek elemente srpskoga ili srpsko-hrvatskoga jezika.
2. Na mnogim internetskim stranicama tekstovi sadržavaju mnogo suvišnih elemenata te se zbog toga prije samoga uvrštavanja u korpus oni moraju pročistiti, odnosno prethodno obraditi.

Razlog odabiranja blogova kao tekstne vrste leži u činjenici da bi korpus trebao sadržavati i elemente govorenoga jezika. Blogovi su izabrani iz razloga što su, prema zaključcima istraživanja Nowsona, Oberalandera, i Gilla (2005), oni klasificirani kao žanr sa službenijim stilom pisanja od e-mailova, ali neslužbenijim od primjerice biografija ili školskih eseja, čime se ujedno odražava i neslužbenost u komunikaciji[86] (Nowson-Oberlander-Gill 2005: 1668). Naime, spomenuta se saznanja temelje na istraživanju u kom su znanstvenici na temelju F-mjerila (*F-measure*), koji je jedino mjerilo neformalnosti (kontekstualnosti) tekstova nasuprot formalnosti, gdje niska vrijednost predstavlja kontekstualnost, veću uporabu zamjenica, glagola, priloga i interjekcija, dok viša vrijednost predstavlja formalnost, a odražava se u većoj uporabi imenica, pridjeva, prijedloga i članova (Nowson-Oberlander-Gill 2005: 1667), određivali F-vrijednosti pojedinih žanrova na temelju korpusa. Blogove su time smjestili na sredinu svoje ljestvice što dokazuje njihovu važnost kao tekstne vrste koja je po svome stilu najbliža razgovornome jeziku. Važno je pri tome napomenuti da je ovo istraživanje provedeno na temelju e-mail i weblog arhiva Britanskoga nacionalnoga korpusa (BNC) te da se samo pretpostavkom rezultati istraživanja mogu primijeniti i na druga govorna područja, poput primjerice hrvatskoga.

[86] Prema izvorniku: „Now, according to the results in the first part of this paper (see Table 2), blogs are more formal than e–mail. But they are still relatively informal—just surpassing School Essays in their F–score. Even ignoring the non–linearity just noted, this relative informality could reflect informality in the communication situation.“

Tekstovi koji su uvršteni u četvrti potkorpus isto tako sadržavaju poseban registar, odnosno jezik stručne literature, dok su u peti testni potkorpus ušli tekstovi iz različitih udžbenika za različite uzraste.

Iz razloga što broj tekstova u korpusu i pojedinim potkorpusima nije ograničen, postoji stalna mogućnost proširivanja testnoga korpusa dodavanjem novih materijala te sukladno tomu i dodavanja novih lema, odnosno poboljšanja trenutne inačice morfološkoga analizatora i morfološkoga rječnika.

7. Primjena morfološkoga analizatora za učenje i poučavanje hrvatskoga standardnog jezika

Kao što se pokazalo u poglavlju 2.3.5., postoji povećana potreba za razvijanjem pomagala kojim će se koristiti u učenju i poučavanju hrvatskoga kao drugoga ili stranoga jezika. Međutim u ovoj se fazi rada postavlja pitanje svrhe izrade jezičnoga pomagala, odnosno definiranje same strukture računalnoga programa kojim će se koristiti za učenje i poučavanje hrvatskoga kao drugoga ili stranoga jezika. Iz razloga što se već prije promišljalo te realizirao projekt izrade Hrvatskoga morfloškog leksikona koji je dostupan preko interneta[87] te koji pruža mnogo korisnih informacija za govornike kojima hrvatski nije materinski jezik, u ovoj se fazi projekta teorijski promišljalo o izradi pomagala koje će imati ili karakteristike morfološkoga leksikona, tezaurusa ili pak morfološkoga rječnika.
Ako se promotre pojedine definicije iz Anićeva rječnika relevantne za spomenuti problem, vidimo da se navedeni pojmovi tumače na sljedeći način:

leksikon – 1. knjiga u kojoj su abecednim redom protumačeni različiti pojmovi, imena, događaji i stručno nazivlje, itd.; 2. a. rječničko blago nekoga jezika, leksik, b. terminologija neke struke; nazivlje
tezaurus – 2. *inform.* a. rječnik riječi svrstanih prema sličnosti ili oprečnosti; b. sakupljena rječnička baza prema nekom načelu, c. program koji sadrži takve datoteke (najčešće dio programa za obradu teksta), računalni rječnik
rječnik – 1. skupljene riječi i sklopovi riječi jednog ili više jezika uz objašnjenje njihove upotrebe i značenja; rječnik koji sadrži natuknice i ekvivalente iz jednoga (dva ili više) jezika

Prodrobnije promatrajući dosadašnje radove i publikacije, prema mojim saznanjima, postoje dva programa koji mogu izvrsno služiti učenju i poučavanju hrvatskoga kao drugoga ili stranoga jezika, a to je osim spomenute internetske verzije Hrvatskoga morfološkog rječnika i već analizirani Gramatički tezaurus hrvatskoga jezika. Ako se uzmu u obzir promišljanja o upotrebi programa unutar nastavne discipline CALL (Computer-Aided Language Learning[88]) (Prószéky 1997b), možemo zaključiti da program mora korisnicima koristiti u dvije svrhe: prva je davati informacije o morfološkim odrednicama pojedine riječi, odnosno informacije o određenoj paradigmi, a druga je svrha ponuditi

[87] http://hml.ffzg.hr/hml/, 18. srpnja 2008.
[88] Učenje jezika uz pomoć računala

relevantne informacije o uporabi određene riječi, eventualno davanjem ekvivalenta iz drugoga ili drugih jezika.
Kada se ove informacije uzmu u obzir, može se zaključiti da bi program, čija se izrada postavila ciljem ovoga rada, trebao dobiti status računalnoga morfološkog rječnika. Zaključci iz prethodnih poglavlja pak nadopunjuju njegov naziv te se time može reći da se radi o računalnome morfološkom rječniku hrvatskoga standardnog jezika.

Ako se pogledaju neki od morfoloških rječnika, može se uočiti da su izrađeni na različitim načelima. Dok se češki morfološki leksikon koristi sličnim kategorijama, onima u HUMOR-u, Hrvatski je morfološki leksikon izrađen na načelu generiranja oblika te njihova provjeravanja na korpusu. Skuomalová (1997), opisujući tretiranje fonoloških kategorija češkoga morfološkog leksikona, o leksikonu navodi sljedeće:

> Postojeći je leksikon bio napravljen za jednostavne C programe koji samo dodaju „nastavke“ na „stemove“. Navodnici u prošloj rečenici znače da se izrazi ne rabe u lingvističkome smislu, nego u tehničkome: stem označuje bilo koji dio riječi koji se ne mijenja za vrijeme deklinacije / konjugacije. Nastavak označuje prave nastavke i po mogućnosti ostatak riječi koji se mijenja[89] (Skuomalová 1997: 1).

Tadić i Fulgosi (2003) pri izradi morfološkoga leksikona koristili su se procesom generiranja, a ne procesom analize te se koriste dvama generatorima oblika, fleksijskim i derivacijskim. Time leksikon obrađuje sve kombinacije morfema prema morfotaktičkim pravilima, a oblici se uspoređuju i provjeravaju na velikome korpusu (Tadić-Fulgosi 2003: 41).

Morfološki rječnik koji je zadan ciljevima ovoga rada temelji se na načelu morfološke analize i korisnicima pruža formalistički pristup koji se u nekim slučajevima kosi s tradicionalnim morfološkim kategorijama. Tehničke karakteristike morfološke analize, prednosti i korisnosti na sastavljanje ovoga rječnika već su pojašnjene u prethodnim poglavljima. Ostaje još samo otvoreno pitanje formalizma i prednosti njegova korištenja nasuprot tradicionalnim rješenjima, što će biti objašnjeno u sljedećem poglavlju. Ipak,

[89] Prijevod MA. Izvornik: “The existing lexicon was originally designed for simple C programs that only attach “endings” to the “stems”. The quotation marks in the previous sentence mean that the terms are not used in the linguistic meaning but rather technically: Stem means any part of a word that is not changed in declension/ conjugation. Ending means the real ending and possibly also another part of the word that is changed.“

samu dvojbu o tome olakšava li ili otežava li formalistički pristup usvajanje gramatike može ipak otkloniti samo ciljano istraživanje.

7.1. Problematika obrade promjenjivih vrsta riječi i primijenjena rješenja

Pri izradi morfološkoga parsera pojavilo se nekoliko pitanja koja su se morala riješiti. U ovome će se poglavlju prikazati najvažnija problematika koja se pojavila pri obradi imenske, pridjevske i glagolske paradigme. Zasigurno se postavlja pitanje zbog čega se izabrala prezentacija samo ovih vrsta riječi, a ne svih promjenjivih vrsta riječi koje su se obradile tijekom prilagođavanja morfološkoga parsera HUMOR-a sustavu hrvatskoga standardnog jezika.

Naime, uzimajući u obzir posebnosti samoga rada i njegov opseg, razmatrat će se samo specifična problematika, odnosno problematika čija rješenja izravno utječu na usvajanje hrvatskoga kao drugoga ili stranoga jezika. Prikazana problematika koja je nastala, smatra se tipičnom problematikom te se zbog toga odlučilo na njezin detaljniji prikaz. Problematika obrade ostalih vrsta riječi koje pripadaju kategoriji promjenjivih vrsta riječi veže se uz rješenja koja su predočena u ovome poglavlju te bi njihov opis predstavljao ponavljanje elementa, što se pokušalo izbjeći.

Pri razmatranju redoslijeda prikazivanja navedenih vrsta riječi odlučilo se ponajprije za prikaz imenske, zatim pridjevske te naposljetku glagolske paradigme, jer se pitanja koja se razmatraju unutar glagolske paradigme izravno nadovezuju na problematiku koja je predstavljena u prethodnim poglavljima.

7.2. Imenice

U rječničkoj osnovici programa postoji 27 866 imenica, što čini okosnicu i morfološkoga rječnika. Ako se pogledaju dosadašnje prezentacije imenske paradigme, možemo uočiti različitost pristupa, što je već pojašnjeno u poglavlju 2.3.3, odnosno nemogućnost izravne primjene opisanih imenskih paradigmi na sustav automatske morfološke analize hrvatskoga standardnog jezika. U ovome će se poglavlju pozornost posvetiti problematici koja se pojavila pri opisivanju imenske paradigme u sustavu HUMOR. Osim navedene

problematike prikazat će se i rješenja koja su se mogla poduzeti, odnosno naznake alternativnih rješenja koja bi trebala stupiti na snagu u nastavku samoga projekta. Važno je napomenuti da se imenicama smatraju sve one riječi koje imaju imensku paradigmu.

Prvo pitanje koje se pojavilo kada se pristupilo opisivanju imenske paradigme u sustavu HUMOR, odnosilo se na problematiku roda i padeža. Naime, osim dva broja, Raguž (1997: 6) primjerice spominje još jedan broj, dvojinu ili dual. Podrobnijom analizom navoda u PHG utvrđeno je da s morfološkoga stajališta „dvojina imenica nema nikakva zasebna oblika u hrvatskome jeziku. (Uvijek je to jedan od oblika koji već postoje u deklinaciji imenica.)“ (Raguž 1997: 6).
Morfološki se parser HUMOR, naime, koristi za analizu pisane varijante hrvatskoga standardnog jezika, što je već pojašnjeno u poglavlju 6.2. Sa stajališta morfološke analize oblici u dvojini u pisanome jeziku osim toga uvijek se podudaraju s već postojećim oblicima, te je zbog toga ovaj problem pri izradi morfološkoga parsera trenutno marginaliziran.
Problematika je roda pri izradi parsera isto tako dodatno razrađena (više o tome u sljedećem poglavlju). Primijenjena su neka rješenja koja mogu utjecati na olakšano usvajanje imenske paradigme kod učenika kojima hrvatski nije materinski jezik.

7.2.1. *Imenska paradigma*

Kao polazište za analizu imenske paradigme poslužile su paradigme imenica *dječak*, *cvijet* i *more*, što je prikazano u prikazu 19.

Jd.

m.r. živo		m.r. neživo	ž.r.	sr.r
N	dječak	cvijet	stolica	more
G	dječaka	cvijeta	stolice	mora
D	dječaku	cvijetu	stolici	moru
A	dječaka	cvijet	stolicu	more
V	dječače	cvijete	stolico	more
L	dječaku	cvijetu	stolici	moru
I	dječakom	cvijetom	stolicom	morem

Mn.

	m.r. živo	m.r. neživo	ž.r.	sr.r
N	dječaci	cvjetovi	stolice	mora
G	dječaka	cvjetova	stolica	mora
D	dječacima	cvjetovima	stolicama	morima
A	dječake	cvjetove	stolice	mora
V	dječaci	cvjetovi	stolice	mora
L	dječacima	cvjetovima	stolicama	morima
I	dječacima	cvjetovima	stolicama	morima

Prikaz 19: Imenske paradigme imenica *dječak, cvijet, stolica, more*

Da bi učenik kojemu hrvatski nije materinski jezik ili pak računalo moglo generirati paradigme navedenih imenica, prema tradicionalnoj gramatici mora poduzeti sljedeće korake:

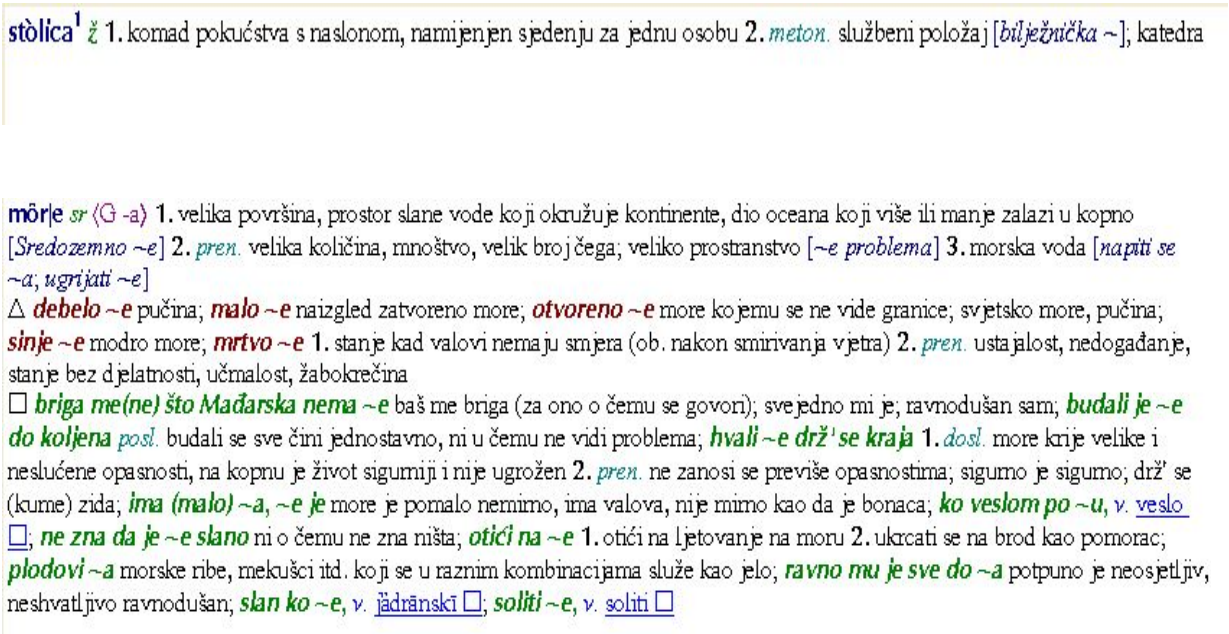
1. Mora znati značenje imenica da bi mogao odrediti radi li se o imenici za živo ili neživo.
2. Mora znati o kojoj se deklinaciji radi *(-a, -e ili –i deklinaciji)* ili mora iz rječnika dobiti informaciju o genitivu jednine navedenih imenica.
3. Mora znati osnovu[90] da bi mogao napraviti određenu paradigmu.

Prva sekundarna literatura koja je učeniku potrebna da bi dobio informacije koje su opisane u prvome koraku, jest rječnik. U drugome koraku učenik još uvijek poseže za rječnikom jer mu trebaju gramatičke informacije da bi odredio o kojoj je deklinaciji riječ. Prikaz 20 sadrži preslike informacija onih lema iz Anićeva rječnika koje su navedene u prethodnoj tablici.

dječāk *m* ⟨G dječáka, V dječāče, N *mn* dječáci⟩ muško dijete poslije prvoga djetinjstva (do puberteta), prije nego što postane mladić

cvijēt *m* ⟨N *mn* cvjētovi, *zb.* cviječe⟩ **1.** *bot.* dio biljke koji sadržava organe za oplodnju i iz kojeg se razvija plod **2. a.** najbolji ili prvi dio čega npr. domaćeg proizvoda (brašna, rakije) **b.** *retor.* najbolji, elita [*~ naroda*]
□ ***u ~u mladosti*** u najljepše doba života; ***život vodenog ~a*** kratak život, kratko trajanje čega

[90] HG navodi da se „nastavci dodaju na osnovu, koja se, u pravilu, dobije ako se u gen. jedn. izostavi nastavak“ (Barić et al. 1995:104)



stòlica[1] ž 1. komad pokućstva s naslonom, namijenjen sjedenju za jednu osobu 2. *meton.* službeni položaj [*bilježnička ~*]; katedra

mȏr|e *sr* ⟨G -a⟩ 1. velika površina, prostor slane vode koji okružuje kontinente, dio oceana koji više ili manje zalazi u kopno [*Sredozemno ~e*] 2. *pren.* velika količina, mnoštvo, velik broj čega; veliko prostranstvo [*~e problema*] 3. morska voda [*napiti se ~a; ugrijati ~e*]
△ ***debelo ~e*** pučina; ***malo ~e*** naizgled zatvoreno more; ***otvoreno ~e*** more kojemu se ne vide granice; svjetsko more, pučina; ***sinje ~e*** modro more; ***mrtvo ~e*** 1. stanje kad valovi nemaju smjera (ob. nakon smirivanja vjetra) 2. *pren.* ustajalost, nedogađanje, stanje bez djelatnosti, učmalost, žabokrečina
□ ***briga me(ne) što Mađarska nema ~e*** baš me briga (za ono o čemu se govori); svejedno mi je; ravnodušan sam; ***budali je ~e do koljena*** *posl.* budali se sve čini jednostavno, ni u čemu ne vidi problema; ***hvali ~e drž' se kraja*** 1. *dosl.* more krije velike i neslućene opasnosti, na kopnu je život sigurniji i nije ugrožen 2. *pren.* ne zanosi se previše opasnostima; sigurno je sigurno; drž' se (kume) zida; ***ima (malo) ~a, ~e je*** more je pomalo nemirno, ima valova, nije mirno kao da je bonaca; ***ko veslom po ~u***, *v.* veslo □; ***ne zna da je ~e slano*** ni o čemu ne zna ništa; ***otići na ~e*** 1. otići na ljetovanje na moru 2. ukrcati se na brod kao pomorac; ***plodovi ~a*** morske ribe, mekušci itd. koji se u raznim kombinacijama služe kao jelo; ***ravno mu je sve do ~a*** potpuno je neosjetljiv, neshvatljivo ravnodušan; ***slan ko ~e***, *v.* jȁdrānskī □; ***soliti ~e***, *v.* soliti □

Prikaz 20: Preslike lema *dječak, cvijet, stolica, more* iz Anićeva rječnika

Kao što je iz ovih prikaza vidljivo, informacije koje su potrebne učeniku ili računalu za određivanje vrste deklinacije dostupne su samo kod lema *more* i *dječak*. Kod imenica *cvijet* i *stolica* ne postoje navodi za genitiv jednine uz pomoć kojih bi učenik mogao zaključiti o kojoj se vrsti deklinacije radi, tako da se mora posegnuti za nekim drugim izvorima. U RHJ kod lema postoje oznake samo za rodove (*m*, *ž*, *sr*, *zb*, *pl. tantum* i *sg. tantum*). Oznake za živo i neživo učenik mora sam semantizirati, što je onemogućeno računalu. Sa stajališta morfološke analize hrvatskoga jezika to znači da bi se ove oznake morale ručno unijeti kod svake pojedine leme.
Za razliku od RHJ, u rječnicima za njemački jezik učenicima su dane sve informacije koje su potrebne za uspješno generiranje paradigme imenica (nastavak za Gjd i Njd).

Pretpostavimo nadalje da je učenik dobio na neki način potrebne informacije te može prijeći na treći korak, odnosno generiranje odgovarajućih paradigmi. Tu se pojavljuje sljedeća dvojba, odnosno potreba dobivanja informacija o osnovi na koju bi se dodali nastavci prema odgovarajućoj paradigmi. U Anićevu rječniku osnova je izričito naznačena samo kod leme *more*. Kod drugih se testnih lema osnova da naslutiti, odnosno neizravno zaključiti ako je znanje hrvatskoga na većem stupnju. Za računalo takvo što dakako nije moguće, jer računalna obrada nekoga jezika zahtijeva eksplicitne naredbe.

Ako se odlučimo za pokusno testiranje primjerice *a-deklinacije* s probnom lemom *dječak*, nailazimo na neke nedoumice. Tablica iz PHG koja bi učeniku kojemu materinski jezik nije hrvatski trebala dati informacije o nastavcima koji se dodaju na osnovu da bi se dobila paradigma zadane imenice, izgleda kao u prikazu 21.

	Muški rod		Srednji rod	
	Jednina	Množina	Jednina	Množina
N	-Ø, -o/e	-i	-o/e	-a
G	-a	-ā, -ī, -ijū	-a	-ā
D	-u	-ima	-u	-ima
A	=N/=G	-e	-o/e	-a
V	-e/u, =N	-i	-o/e	-a
L	-u	-ima	-u	-ima
I	-om/em	-ima	-om/em	-ima

Prikaz 21. Uzorak *a-deklinacije* iz PHG

Prvo pitanje koje se ovdje pojavljuje, odnosi se na problem osnove. Naime, iz razloga što u vokativu jednine dolazi do alternacije osnove, zaključujemo da su za generiranje paradigme potrebne dodatne informacije. Isto tako iz same ove tablice nije vidljivo koji bi se oblik trebao odabrati u akuzativu, a koji u vokativu odnosno instrumentalu jednine te genitivu množine.

Ako bismo pak računalu pokušali objasniti pravilo alternacije primjerice ove osnovice, odnosno kada *k* prelazi u *č*, morali bismo napisati složeni algoritam te u njega uklopiti sve moguće iznimke, što se pokazalo prilično nepraktičnim.

Jedno od rješenja koje je primijenjeno u morfološkome parseru HUMOR-u, a koje bi se moglo pokazati učinkovitim i kod učenja i poučavanja hrvatskoga kao stranoga jezika, jest podjela lema na stemove i termove. Ako bismo tim načelom pogledali neke od lema koje priapdaju *a-deklinaciji*, mogli bismo uočiti da se radi o različitim termovima (Prikaz 22).

LEMA	STEM	TERM
dio	di-	-o
bistro	bistro-	-Ø
pitalac	pita-	-lac
bolero	boler-	-o
intelektualac	intelektual-	-ac
siroče	siroče-	- Ø
ušće	ušć-	-e

Prikaz 22: Podjela nekih lema *a-deklinacije* na stemove i termove

Iz gornjega je prikaza vidljivo da je pri računalnoj morfološkoj analizi hrvatskoga jezika potrebna druga vrsta kategorizacije, odnosno podjela na imenske fleksijske skupine koje bi se naznačile kod svake pojedine riječi.

Primjenom ovoga načina došlo se do 189 imenskih skupina. Iako se ovaj broj možda čini prevelikim kod učenja hrvatskoga kao drugoga ili stranoga jezika, razrađivanje skupina koje se razlikuju po najmanje jednom termu od neke druge skupine dalo je povoda za razvijanje sustava informacija koje je potrebno naznačiti uz lemu da bi se učenicima stranih jezika omogućilo pravilno generiranje cjelokupne paradigme.

U prikazu 23 dan je izvadak iz popisa skupina s termovima i primjerima lema za svaku pojedinu skupinu. Važno je pri tome napomenuti da ovdje u prvome planu nisu imenice, nego imenska skupina. Stoga se može dogoditi da neka imenica koja je svrstana kao primjer za imensku skupinu ima još neke druge oblike koji se svrstavaju pod drugu skupinu jer pripadaju nekoj drugoj paradigmi. Imenice navedene kao primjer služe samo za predočavanje određene fleksijske skupine. Kod pojedinih skupina isto je tako moguće da je navedeni primjer jedina imenica koja ima navedenu paradigmu.

Stem je od terma u prikazu odvojen okomitom crtom. Svi se navedeni oblici temelje na informacijama iz HG, RHJ i paradigmama iz GT.

SKUPINA	**TERM**								**PRIMJER**
	11	**12**	**13**	**14**	**15**	**17**	**12**	**22**	
97	o	a	u	o	o	om	a	a	neb|o
98	o	a	u	o	o	om	esa	esa	neb|o
99	e	a	u	e	e	em	a	a	tališt|e
100	a	a	u	a	a	om	a	a	dob|a
101	e	ena	enu	e	e	enom	ena	ena	sjem|e
102	e	eta	etu	e	e	etom	ad	adi	pač|e
103	e	eva	evu	e	e	evom	eva	eva	podn|e
105	lo	la	u	lo	lo	lom	la	ala	z|lo
107	lo	la	lu	lo	lo	lom	lesa	lesa	tije|lo
108	mo	ma	mu	mo	mo	mom	ma	ama	pis|mo
109	no	na	nu	no	no	nom	ni	ana	d|no
110	no	na	nu	no	no	nom	ni	na	steg|no
111	ro	ra	ru	ro	ro	rom	ra	ara	reb|ro
112	pko	pka	pku	pko	pko	pkom	pka	baka	klu|pko
113	pko	pka	pku	pko	pko	pkom	pka	pka	klu|pko

Prikaz 23: Izvadak iz tablice s imenskim skupinama koje se koriste u morfološkome parseru HUMOR-u

Ako se podrobnije promotri gornji prikaz, mogu se uočiti neke univerzalije koje se tiču generiranja oblika u pojedinim paradigmama. Cilj je ovoga formalističkog pristupa

generiranje paradigme s točnim oblicima u svim padežima. Ako su kod nekih imenica navedenih u stupcu s primjerima mogući i dvojni oblici, oni se svrstavaju pod zasebnu skupinu. Primjerice, prema GT imenica *pače* ima još i oblik u instrumentalu množine *pačađu* koji nije naveden u ovome izvatku jer pripada skupini koja ovdje nije prikazana, dok se 97. i 98. te 112. i 113. skupina razlikuju u svega dva ili samo jednome obliku.

Analizirajući navedenu podjelu po skupinama, može se uočiti da je učeniku potrebno nekoliko gramatičkih informacija da bi uspješno generirao paradigmu zadane riječi.

1. Učenik mora znati stem riječi.
2. Učenik mora dobiti „ključne“ informacije:
    a) označuje li imenica živo ili neživo,
    b) kako glase termovi 11 (Njd), 21 (Gjd), 12 (Nmn), 22 (Gmn).

Kada bi se leme u morfološkome rječniku nadopunile ovim informacijama korištenjem matrice[91] kao u prikazu 24, dobila bi se gotovo cijela paradigma zadane riječi.

Jd.

N	11
G	21
D/L	21-[1]+ ♠♣u/♦i
A	♠♣ #11/*12 ♦21-[1]+u ♦21\i = 11
I	[1]♦♠♣21-[1]+om [2]♠♣21-[1]\c;j+em ♦21\i = 21

Mn.

N/V/♣A	12
G	22
D/L/I	12-[1]+ ♦ama/♠♣ima ♦12\i + ma
A♦♠	[1]12-[1]+e [2]12-[1]\c-[1]+ke ♦12\i = 12

♠ m.r, ♣sr.r; ♦ž.r.
živo, * neživo
\ „ako je“
[1] jedno slovo
= onda je
; „ili“

Prikaz 24. Matrica za generiranje imenske paradigme

Rječnički unos primjerice za lemu *more* izgledao bi kao u primjeru (1), a za lemu *dječak* izgledao bi kao u primjeru (2)

[91] Sustav ove matrice i računalnih rodova pokriva većinu imenica iz Anićeva rječnika. Iznimke koje se nekim slučajem ne uklapaju u navedenu matricu, bit će navedene zasebno. Vjd. namjerno je izostavljen iz paradigme. Više o problematici određivanja vokativa u sljedećem odlomku.

(1) mor|e sr.r (nž., -e,-a,-a,-a)
(2) dječa|k m.r (ž.,-k,-ka,-ci,-ka)

Važno je pri tome napomenuti da u morfološkome parseru veliku ulogu imaju i računalne oznake za rodove, o čemu će biti govora u sljedećem poglavlju. Oznake za rodove određuju paradigmu. Tako je primjerice kod imenice *dijete* oznakom za rod određena paradigma u jednini, koja se isto povezuje s oblikom *djec|a*….
Rječnički unos tada bi za lemu *dijete* izgledao kao u primjeru (3)

(3) d|ijete sr.r (ž. –ijete, jeteta, *v. djec|a*)
djec|a sr.r.1 (ž. *v. d|ijete,* -a, -e)[92]

Rječničkim unosima u morfološkome se rječniku dodaju još i oznake za imenske skupine koje omogućuju jednoznačno korištenje gornje matrice. Sve u svemu, učinkovitost ovoga sustava provjerit će se pokusnim istraživanjem u nastavku ovoga projekta.

7.2.2. *Problematika*

Računalni morfološki opis imenske paradigme nastao je na načelu rada morfloškoga parsera HUMOR-a. Kao što se u prethodnom odlomku moglo vidjeti, opis je rezultirao nekim rješenjima koja se mogu primijeniti i kod učenja odnosno poučavanja hrvatskoga kao drugoga ili stranoga jezika.
Međutim, iako su se postavile imenske kategorije od kojih se svaka sastoji od termova koji odgovaraju padežima u jednini i množini, pojavili su se problemi od kojih će se neki u ovome odlomku samo natuknuti, a detaljno objasniti u sljedećim poglavljima.

Jedan od važnih problema koji su se pojavili prvenstveno kod imenske paradigme jest problem određivanja nekih padežnih oblika. Kao što će biti objašnjeno, sa stajališta učenja jezika i računalnoga morfološkog opisa hrvatskoga standardnog jezika smatra se učinkovitijim padeže dativ i lokativ jednine svrstati pod istu stavku. Naime, na morfološkoj

[92] Ovaj se rod dodaje u prethodno navedenu tablicu pokraj oznake za ž.r. Objašnjenje zašto je to tako, slijedi u sljedećem poglavlju. U prikazu 24 on je izostavljen da bi se izbjegla pomutnja.

se razini do sada nije uočio nijedan primjer u kojemu bi se oblici u dativu i lokativu razlikovali. Njihovim svrstavanjem pod istu stavku govornicima kojima hrvatski nije materinski jezik uvelike se olakšava usvajanje sklonidbene paradigme.
Drugi problem koji se pojavio, odnosio se na problem vokativa jednine. Mnogi su jezikoslovci mišljenja da vokativ nije potreban jer se rijetko upotrebljavaju (dozivaju) primjerice imenice koje označavaju stvari (više o tome u poglavlju 8.4.1.). Sa stajališta računalne obrade jezika potrebne su ove informacije jer se svakoj pojedinoj lemi mora pridružiti oblik i u vokativu. Njegovu stvarnu upotrebu pokazat će korpusna analiza.

Kada se pokušalo naći odgovore na postavljena pitanja koja su se odnosila na oblike u vokativu jednine pojedinih imenica, došlo se do zaključka da jezični priručnici ne pokrivaju sve slučajeve. Iako primjerice u gramatikama stoji da su u vokativu jednine mogući nastavci *–e* i *–u* te u nekim slučajevima i oba nastavka, nigdje se ne preciziraju oni slučajevi kada su obje inačice moguće. Dvojbenim su se imenicama pokazale imenice *tumor*, *tenor*, *gospodar*, *cipelar, novinar*, *Alah* i *Vlah* te imenice koje označuju neživo, primjerice *krug, dug, lug, jug,* i *rug.* Odgovore na postavljena pitanja ne može dati ni korpusna analiza jer se u mnogo slučajeva oblici u vokativu podudaraju s već postojećim oblicima u paradigmi.

Sljedeća problematika koja se pojavila odnosila se na automatsku obradu pojedinih rječničkih unosa. Pojavili su se naime slučajevi za koje se pretpostavljalo da imaju iste termove, poput imenica *hvat, zahvat, procvat, vrat* i *navrat.* Ako se pogleda paradigma ovih imenica u množini, uočava se da se one razlikuju te dobivaju različite termove u množini kao što je vidljivo u primjerima (4), (5), (6), (7) i (8).

(4) *hvat- hvat+ovi,*
(5) *zahvat-zahvat+i,*
(6) *procvat- procvat+i,*
(7) *vrat-vrat+ovi,*
(8) *navrat-navrat+i*.

Iz razloga što u nekim slučajevima dolazi do promjene samoga stema, povećava se i broj termova te se naoko srodne riječi svrstavaju pod različite imenske skupine:

(9) *svijet- sv+jetovi,*

(10) *probisvijet- probisvijet+i.*

Osim navedenih primjera, problematika određivanja paradigme pojavila se i kod pojedinih lema koje se nalaze u Anićevu rječniku, a koje se koriste regionalno ili lokalno, poput imenica *rsuz, hrsuz, rkać* i *broć*. Rijetkost ovih natuknica dokazuje i činjenica da se osim imenice *broć* koja se pojavljuje jedanput, ostale tri imenice ne pojavljuju nijednom u Hrvatskome nacionalnom korpusu koji sadrži 101,3 milijuna pojavnica.

Navedeni primjeri ukazuju na činjenicu da se opis imenske paradigme u računalnoj morfološkoj analizi hrvatskoga standardnog jezika ne može obaviti automatski, nego poluautomatski, odnosno ručnom provjerom i ispravkom navoda.

7.3. Pridjevi

Jedna od promjenjivih vrsta riječi koja stvara dodatne poteškoće pri učenju hrvatskoga kao drugoga ili stranoga jezika jesu pridjevi[93]. Ako se pogledaju dosadašnji opisi pridjevske paradigme predočeni u poglavlju 2.3.2., može se uočiti da različiti autori na različiti način objašnjavanju deklinaciju pridjeva. Sa stajališta učenika hrvatskoga kao drugoga ili stranoga jezika ovi se opisi mogu učiniti teškima, a cijela paradigma gotovo nesvladivom.
U ovome poglavlju slijedi prikaz formalističkoga pristupa pridjevskoj paradigmi koji je nastao unutar izrade morfološkoga analizatora HUMOR-a. Rješenja koja su primijenjena pri izradi parsera mogu se primijeniti i na učenje i poučavanje hrvatskoga kao drugoga ili stranoga jezika, te olakšati usvajanje ovoga gradiva.

Prije pojašnjavanja pridjevske paradigme potrebno je definirati i objasniti sljedeće stavke:

1. Iz razloga što pridjevsku paradigmu unutar izrade morfološkoga parsera mogu imati ne samo pridjevi, nego i neke druge vrste riječi (primjerice zamjenice), pod pojmom pridjevi ne smatraju se samo pridjevi kao vrsta riječi, nego sve one vrste riječi koje se svojom paradigmom uklapaju u pridjevsku paradigmu.

[93] Izjava se temelji na internoj anketi provedenoj među nastavnicima koji podučavaju hrvatski jezik govornicima kojima hrvatski nije materinski jezik.

2. Cilj je ovoga dijela rada prikaz rješenja koja su nastala uslijed izrade morfološkoga analizatora, a koja se mogu izravno primijeniti na učenje i poučavanje hrvatskoga kao drugoga ili stranoga jezika. Stoga će se usredotočiti samo na one pridjeve koji imaju različite oblike u svojoj paradigmi. Nepromjenjivim pridjevima poput primjerice *super, top* i *fit*, a koji su navedeni u rječničkoj osnovici programa odnosno u Anićevu rječniku, neće se pridavati veća pozornost.

Kada bismo morali sažeti dosadašnje analize pridjevske paradigme u rječnicima i priručnicima dostupnima za učenje i poučavanje hrvatskoga jezika iz poglavlja 2.3.2, mogli bismo uočiti da više autora navodi različite kategorije pridjeva. Učenike mnoštvo ovih opisa može zbuniti. GHJ dijeli pridjeve prema njihovim semantičkim osobinama te ih svrstava pod kategorije *kakvoćnih (kvalitativnih) pridjeva, posvojnih (posesivnih) pridjeva, gradivnih (materijalnih) pridjeva* i *odnosnih pridjeva* (2005: 133f) koji se ne spominju u PHG. Ako se pak pogledaju kategorije koje navodi PHG (Raguž 1997: 88), vidimo da on spominje kategorije *opisnih, odnosnih* i *posvojnih pridjeva*. HG navodi kod podjele pridjeva da se pridjevi po značenju mogu podijeliti na samo dvije skupine, na *opisne* i *odnosne*, s tim da *odnosni pridjevi* obuhvaćaju *posvojne* i *gradivne*.

Ako se pogledaju pridjevske paradigme prema ovim opisima, uočava se mnoštvo različitih pristupa pridjevskoj deklinaciji (poglavlje 2.3.2.). Kada se tomu doda još i činjenica da se ovoj podjeli pridjeva u pravilu dodaje još i pridjevski vid koji se povezuje sa semantičkim svojstvima pridjeva jer se navodi da *određeni* i *neodređeni vid* imaju samo opisni pridjevi (HG 1995: 179), dolazimo do zaključka da je na osnovi ovih kategorija i podjela gotovo nemoguće napraviti morfološki parser.

Ako pak na osnovi ovih informacija promotrimo korake koje učenik hrvatskoga kao drugoga ili stranoga jezika ili pak računalo prije generiranja pridjevske paradigme ili traženja nekoga pridjevskog oblika mora poduzeti, dolazimo do sljedećih zaključaka:

1. Učenik (ili računalo) mora u prvome koraku znati značenje samoga pridjeva da bi ga mogao svrstati pod jednu od kategorija.
2. Ako je pridjev opisni, učenik (ili računalo) mora odrediti i njegov vid.
3. Učenik (ili računalo) mora znati koja imenica slijedi da bi mogao odrediti radi li se o deklinaciji za živo ili neživo.

Ako se na osnovi tradicionalnih opisa nadalje promotri pridjevska paradigma pridjeva *crven*, koja se koristi kod analize pisanoga teksta, vidimo da se ona sastoji od 93 deklinacijska oblika koje učenik mora zapamtiti, odnosno od ukupno 123 različita deklinacijska oblika (Prikaz 25). Ako se u obzir uzme i naglasak, broj se oblika u tradicionalnoj paradigmi povećava.

Jednina:

	m.r., određeni živo	m.r., određeni neživo	m.r., neodređeni živo	m.r., neodređeni neživo	ž.r., određeni	ž.r., neodređeni	sr.r., određeni	sr.r., neodređeni
N	crveni	crveni	crven	crven	crvena	crvena	crveno	crveno
G	crvenog/ga	crvenog/ga	crvena	crvena	crvene	crvene	crvenog/ga	crvena
D	crvenom/me/mu	crvenom/me/mu	crvenu	crvenu	crvenoj	crvenoj	crvenom/me/mu	crvenu
A	crvenog/ga	crveni	crvena	crven	crvenu	crvenu	crveno	crveno
V	crveni	crveni	-	-	crvena	-	crveno	-
L	crvenom/me/mu	crvenom/me/mu	crvenu	crvenu	cvenoj	crvenoj	crvenom/me/mu	crvenu
I	crvenim	crvenim	crvenim	crvenim	cvenom	crvenom	crvenim	crvenim

Množina:

	m.r., određeni	m.r., neodređeni	ž.r., određeni	ž.r., neodređeni	sr.r., određeni	sr.r., neodređeni
N	crveni	crveni	crvene	crvene	crvena	crvena
G	crvenih	crvenih	crvenih	crvenih	crvenih	crvenih
D	crvenim/ma	crvenim/ma	crvenim	crvenim	crvenim/ma	crvenim/ma
A	crvene	crvene	crvene	crvene	crvena	crvena
V	crveni	-	crvene	-	crvena	-
L	crvenim/ma	crvenim/ma	crvenim/ma	crvenim/ma	crvenim/ma	crvenim/ma
I	crvenim/ma	crvenim/ma	crvenim/ma	crvenim/ma	crvenim/ma	crvenim/ma

Prikaz 25: Paradigma pridjeva *crven*

Sa stajališta računalne obrade hrvatskoga jezika bilo je potrebno navedenu pridjevsku paradigmu formalizirati da bi se ona na što jednostavniji način uklopila u obrazac koji računalno može „razumjeti“. Pri tome se pojavilo nekoliko problema.

Prvi se problem odnosio, kao što je već natuknuto, na nemogućnost semantičke kategorizacije pridjeva jer te oznake nisu navedene u Anićevu rječniku. Sve bi se navedene semantičke kategorije morale dakle ručno određivati, što povećava mogućnost pogrješke.

Isto je tako učeniku hrvatskoga kao stranoga jezika onemogućeno generiranje paradigme bez semantičke pozadine[94].

Sukladno načinu rada HUMOR-a rješenje se nametnulo u obliku razdvajanja oblika na stemove i termove. Ako se pogledaju navedeni oblici, uočava se da pridjevska paradigma sadrži manje različitih oblika. Iako mnogo oblika ima svoju dugu i kratku inačicu te se zbog toga broj termova u računalnome morfološkom opisu hrvatskoga jezika povećava, prethodna se paradigma može svesti na ukupno 16 različitih oblika. Oblik 16 u slučaju pridjeva *crven* podudara se s oblikom 8, ali kod nekih pridjeva on je različit te je ovdje zasebno naveden. Primjerice, kod pridjeva *krnj* oblik broj 8 glasi *krnjim*, dok je oblik pod brojem 16 *krnjom*.

Ovih 16 oblika koji se rabe kod morfološkoga parsera HUMOR-a prikazano je na primjeru oblika pridjeva *crven* u sljedećem prikazu:

1 crven-Ø
2 crven-a
3 crven-u
4 crven-im
5 crven-i
6 crven-og
7 crven-oga
8 crven-om
9 crven-ome
10 crven-omu
11 crven-e
12 crven-ih
13 crven-ima
14 crven-oj
15 crven-o
16 crven-om

Prikaz 26: 16 oblika pridjeva *crven* koji se koriste u HUMOR-u

Ako se ovi oblici koje koristi morfološki parser uvrste u jednu matricu, dobiva se tablica koja odgovara prethodnoj paradigmi (Prikaz 27).

[94] Generiranje i analiza oblika riječi bez uporabe semantičke pozadine jedna je od temeljnih odrednica koje je postavila računalna lingvistika pri morfološkoj analizi jezika. Iako je teško pretpostaviti da će učenik jednoga jezika ikada biti u prilici generirati paradigmu ili određene oblike bez semantičkoga znanja, prikaz pridjevske paradigme koja ne zahtijeva semantičku kategorizaciju uvelike olakšava i ubrzava njezino učenje i poučavanje.

Jd.	m.r., neživo	m.r., živo	s.r.	ž.r.
N/#V	*1 / # 5		15	2
G	*2 / #6,7			11
D	*3 / #8,9,10			14
A	*1 / #5	*2 / #6,7	15	3
L	*3 / #8,9,10			14
I	4			16

Mn.	m.r.	s.r.	ž.r.
N/#V	5	2	11
G	12		
D	4,13		
A	11	2	11
L	4,13		
I	4,13		

*** neodređeni, # određeni**

Prikaz 27: Tradicionalna paradigma sa 16 oblika koji se koriste u HUMOR-u

Kada bismo navedene oblike sveli na oblike sa stemovima i termovima, dobili bismo sljedeći prikaz. Oblik broj 16 određen je pridjevskom skupinom (Prikaz 28), što znači da se kod svake pridjevske kategorije definira koji oblik broja 16 ti pridjevi dobivaju (oblik STEM+om ili STEM+em)

1	STEM+Ø
2	STEM+a
3	STEM+u
4	STEM+im
5	STEM+i
6	STEM+og
7	STEM+oga
8	STEM+om
9	STEM+ome
10	STEM+omu
11	STEM+e
12	STEM+ih
13	STEM+ima
14	STEM+oj
15	STEM+o
16	STEM+om/em

Prikaz 28. Normirana tablica sa 16 oblika koji se koriste pri generiranju pridjevske paradigme

Ako se uzme u obzir da se kod pridjevske paradigme sa stajališta računalnoga morfološkog opisa hrvatskoga jezika može uočiti da su oblici u dativu i lokativu uvijek isti, odnosno da im se u množini pridružuje i instrumental, tablica iz prikaza 27 može se još više sažeti, pri čemu se uvelike olakšava učenje i poučavanje pridjevske paradigme u hrvatskome jeziku.

Jd.	m.r., neživo	m.r., živo	s.r.	ž.r.
N/#V	*1 / # 5		15	2
G	*2 / #6,7			11
D/L	*3 / #8,9,10			14
A	*1 / #5	*2 / #6,7	15	3
I	4			16

Mn.	m.r.	s.r.	ž.r.
N/#V	5	2	11
G	12		
D/L/I	4,13		
A	11	2	11

*** neodređeni, # određeni**

Prikaz 29: Pridjevska paradigma sa 16 oblika za učenje i poučavanje hrvatskoga kao drugoga ili stranoga jezika

U rječničkoj osnovici morfološkoga parsera HUMOR-a postoji 9850 pridjeva. Broj pridjeva razlikuje se od broja pridjeva koji se nalaze u Anićevu rječniku. Neke su leme izbačene iz razloga što se ne smatra potrebnim odnositi se prema njima kao prema zasebnim rječničkim unosima jer su prema tvorbi riječi primjerice imenski derivati poput pridjeva *tatin* (ovaj se pridjev u Anićevu rječniku može pronaći kao zasebna lema).

Prema stemovima i termovima svi se pridjevi iz rječničke osnovice mogu podijeliti u 21 pridjevsku (deklinacijsku) skupinu (Prikaz 30). Pri izradi morfološkoga rječnika uz svaku će se lemu navesti i fleksijska kategorija, u ovome slučaju pridjevska skupina. Time se konkretiziraju deklinacijske paradigme riječi i olakšava usvajanje jezika.

Sukladno načelu rada morfološkoga parsera HUMOR-a, kao što je pojašnjeno u poglavlju 4.1., svaki oblik može se sastojati od Ø-terma, ali ne može imati Ø-stem.

pridj. skupina	*term* pridjeva u N.jd.		primjer	broj pridjeva koji pripadaju skupini
	oblik br. 1	oblik br. 5		
I	Ø	-i	crven	2185
II	-	-i	mačji	654
III	-	-i	hrvatski	4879
IV	-ar	-ri	dobar	14
V	-ao	-li	zao	34
VI	-o	-li	debeo	41
VII	-tao	-li	odrastao	2
VIII	-an-	-ni	čudan	1894
IX	-žak	-ški	težak	2
X	-alj	-lji	šupalj	1
XI	-ak	-ki	plitak	24
XII	-tan	-ni	koristan	47
XIII	-zak	-ski	uzak	9
XIV	-dak	-tki	sladak	7
XV	-o	-jeli	cio	1
XVI	-io	-jela	ishlapio	4

XVII	-av	-vi	ovakav	1
XVIII	Ø	-i	krnji	2
XIX	-bak	-pki	gibak	3
XX	-al	-li	obal	7
XXI	-ben	-bni	dioben	1
indekl.	pridjevi koji se ne sklanjaju			38

Prikaz 30: Pridjevske skupine prema termovima s brojem pridjeva koji pripadaju pojedinoj skupini iz rječničke osnovice HUMOR-a

Osim navedenih kategorija pri računalnoj obradi pridjevske paradigme ovim se skupinama dodaje i oznaka iz fleksijske matrice kojom se dodatno usavršava paradigma pojedinih oblika riječi. To se posebice odnosi na primjerice posvojne pridjeve[95] koji se svojim oblicima uklapaju u jednu od skupina, ali nemaju određenu deklinaciju, kao i na one opisne pridjeve koji nemaju neodređeni oblik (primjerice kod pridjeva *muskulaturni* u Anićevu rječniku nije naveden njegov neodređeni parnjak).

Korištenjem ovih šesnaest oblika i pridjevskih skupina isključuje se semantička kategorizacija pridjeva. Kao što se iz gornjega prikaza može iščitati, najvažnijima su se pri određivanju skupina pokazali oblici broj 1 i 5. Ako se podrobnije promotri tablica, isto se tako može vidjeti da svi pridjevi koji čine rječničku osnovicu morfološkoga parsera te su navedeni u Anićevu rječniku, imaju oblik broj 5. Iz razloga što pojedine skupine nemaju neodređeni oblik, valjalo bi možda razmisliti o mogućnosti uvrštenja neodređenoga oblika kao rječničkoga oblika pridjeva u rječnicima koji služe učenju hrvatskoga kao drugoga ili stranoga jezika.

7.3.1. *Stupnjevanje pridjeva*

Jedan od problema koji se pojavio pri izradi računalne pridjevske paradigme bilo je stupnjevanje pridjeva. Iz razloga što se komparativ pridjeva u hrvatskome jeziku tvori uz pomoć sufiksa, primjena tradicionalnih paradigmi ne bi bila moguća jer bi se na taj način uvelike povećao broj termova kod svake pojedine pridjevske skupine. Zbog toga se odlučilo povećati rječničku osnovicu programa stemovima u komparativu i superlativu.

[95] Posvojni pridjevi ne čine ovdje zasebnu skupinu jer nisu posebno naznačeni u Anićevu rječniku koji čini osnovicu rječničke baze morfološkoga analizatora.

Time se zadržava onih 16 oblika kod pridjevske paradigme. U računalnome su opisu komparativni i superlativni stemovi povezani s polazišnim lemama te se na taj način ne sputava ili mijenja rad morfološkoga parsera.

7.3.2. *Jezične nedoumice*

Pri računalnome opisu pridjevske paradigme pojavilo se nekoliko nedoumica. Jedna je takva nedoumica deklinacija pridjeva koji prema Anićevu rječniku mogu imati samo ženski rod, poput primjerice pridjeva *trudna* (u značenju „koja nosi dijete u utrobi"), *muškobanjasta* (značenje: „koja je poput muškarca") ili *suprasna* i *skotna*. Prema navodima u rječniku, deklinacija ovih pridjeva izuzima oblike u muškom i srednjem rodu. Problem se pojavljuje kada jezična uporaba dokazuje mogućnost, odnosno postojanje oblika i u muškome rodu, kao što je to slučaj s pridjevom *trudna*. U hrvatskome prijevodu Hašekova (2004) djela stoji naime navod:

(11) „A jedan je gospon tam bil trudan i zval je svakog na krstitke." (Hašek 2004: 37),

Iz primjera (11) vidljivo je da se pridjev *trudan* koristi u prethodno navedenome značenju „koja nosi dijete u utrobi" u muškome rodu[96], što je proturječno navodima da ovaj i slični pridjevi mogu imati paradigmu samo u ženskom rodu. Naime, jedna od vodilja pri izradi morfološkoga parsera HUMOR-a odnosila se na pravilo generiranja svih mogućih oblika. Morfološkom će se analizom tekstova time omogućiti prepoznavanje i onih oblika koji se poput oblika pridjeva *trudna* pojavljuju u jezičnoj uporabi, a nisu predviđeni i jezičnim priručnicima.

Drugi problem koji se pojavljuje odnosi se na kategoriju priloga i njihovo svrstavanje pod neku od paradigmi. PHG tako spominje podrijetlo priloga i pridjevskim prilozima nastalima od opisnih i odnosnih pridjeva (Raguž 1997: 270-271). U računalnome

[96] Pridjev *trudna* uzet je samo kao primjer navedene problematike. Problem je paradigme ovoga primjera riješen, jer naime postoji lema *trudan* u značenju „koji je obuzet umorom, utrđen, umoran" te se time neizravno riješila paradigma i morfološka analiza oblika ovoga pridjeva u svim rodovima. Kako se iz morfološkoga parsinga izuzima semantika, a u obzir se uzimaju samo oblici u pisanome standardnome hrvatskom jeziku, u morfološki su opis iz navedenoga razloga uvršteni svi oblici.

morfološkom opisu hrvatskoga jezika prilozi nastali od pridjeva svrstani su pod pridjevsku paradigmu jer im se oblik podudara s pridjevskim oblikom broj 15.
Kako je pri računalnome morfološkom opisu hrvatskoga jezika trebalo uzeti u obzir i činjenicu da se prema gramatičkim svojstvima i prilozi mogu stupnjevati, postavila se potreba za rješavanjem ove nedoumice. Iz razloga što je na početku ovoga poglavlja definirana kategorija pridjeva kao svih onih riječi koje imaju pridjevsku paradigmu, nije se smatralo potrebnim uvoditi zasebnu obradu ove vrste riječi. Jedan od problema koji su se pri tome pojavili odnosio se na stupnjevanje pridjeva II. i III. skupine. Raguž (1997: 271) naime tvrdi da od pridjeva koji završavaju na *–i*, samo pridjevi na *–ski* i *–čki* mogu tvoriti priloge, koji se mogu nadalje i stupnjevati. Kao što je kod primjera s pridjevom *trudna* natuknuto, i ovdje postoji bojazan pojave nekih „nepredviđenih“ oblika u korpusu te se pri morfološkome opisu primijenilo rješenje kao u spomenutom primjeru.

Treći veliki problem koji se pojavio pri rješavanju problema pridjevske paradigme odnosio se na pravopisne dvojnosti koji nisu navedeni u Anićevu rječniku. Primjerice, prema Babić-Finka-Moguševu (1995) *Hrvatskome pravopisu* postoje dva načina pisanja pridjeva *mađarski – mađarski* i *madžarski* (Babić-Finka-Moguš 1995: 277). U Anićevu rječniku navedena je samo varijanta *mađarski*, koja je i ušla u rječničku osnovicu parsera, a potom i u računalni morfološki rječnik. I sama je korpusna analiza pokazala prisutnost i druge inačice te je ona naknadno uvrštena u rječničku osnovicu programa. Iz razloga što nije moguće procijeniti broj pravopisnih parnjaka u hrvatskome jeziku, nakon obavljene korpusne analize rječnička će osnovica programa biti po potrebi proširena.

7.3.3. *Pridjevska paradigma u računalnoj obradi hrvatskoga jezika u usporedbi s pridjevskim paradigmama u drugim jezicima*

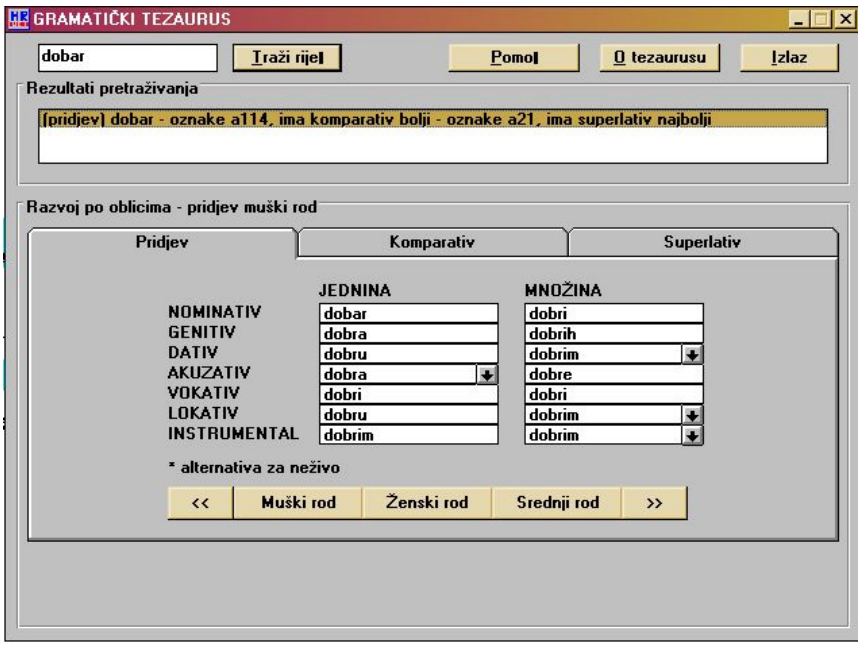
Ako se pogledaju pridjevske paradigme u ostalim jezicima, uočava se težnja za jezičnim formalizmom, odnosno izrada sveobuhvatnih gramatičkih rječnika s određenim fleksijskim kategorijama koje se pridružuju svakoj pojedinoj riječi. U prikazu 31 predočen je izvadak iz Zaliznyjakova (1987) rječnika s naznačenim pridjevom *беспечный* ’bezbrižan’ i pripadajućom fleksijskom skupinom.

бесперебойный 9 п 1*а
бесперемённый 9 п 1*а
бесперемённый 9 п 1*а
бесперечь 2 н
бесперспективность 12 ж 8а
бесперспективный 12 п 1*а
бесперый 5,5 п 1а
беспечальный 7 п 1*а
беспечность 5 ж 8а
беспечный 5 п 1*а
беспилотный 7 п 1*а
бесписьменный 5 п 1*а["1"]
беспламенный 6 п 1*а
бесплановость 6 ж 8а
бесплановый 6 п 1*а
бесплатность 6 ж 8а
бесплатный 6 п 1*а
бесплацкартный 9 п 1*а
бесплодие 6 с 7а

Prikaz 31: Izvadak iz Zaliznyjakova (1987) gramatičkoga rječnika ruskoga jezika *Грамматический словарь русского языка*

Slična se težnja zapazila i kod poljskoga jezika i poljskoga gramatičkog rječnika opisanog u poglavlju 9.1.1.

Ako se pak promotre dosadašnja hrvatska dostignuća, uočava se da jedino GT nudi fleksijske skupine koje na žalost nisu izričito navedene kod prikaza same paradigme (Prikaz 32):



Prikaz 32: Paradigma pridjeva *dobar* iz Gramatičkoga tezaurusa

Kao zaključak prikaza računalne obrade pridjevske paradigme u hrvatskome jeziku nameće se ideja o korištenju navedenih kategorija i skupina ne samo za potrebe morfološke analize hrvatskoga standardnog jezika i izrade računalnoga morfološkog rječnika, nego i u tiskanome obliku za potrebe izrade dodatnih materijala za učenje i poučavanje hrvatskoga kao drugoga ili stranoga jezika. Pokusnim istraživanjem mogla bi se istražiti korisnost ovakvoga pristupa pridjevskoj paradigmi.

7.4. Glagoli

Na temelju analize prikaza glagolske paradigme u jezičnim priručnicima koji se koriste u nastavi hrvatskoga kao drugoga ili stranoga jezika, a koja je pojašnjena u poglavlju 2.3.1., možemo zaključiti da je glagolska paradigma u hrvatskome jeziku jedna od paradigmi kojoj se posvećuje najviše pozornosti tijekom učenja i poučavanja jezika. Raznolikost pristupa i jezičnih objašnjenja ukazuje na činjenicu da bi bilo poželjno osmisliti novi pristup glagolskoj paradigmi koji bi dao odgovore na pitanja učenika te obuhvaćao sve glagole u hrvatskome standardnom jeziku.
U ovome poglavlju slijedi prikaz jednoga takvog pristupa koji je nastao unutar morfološkoga analizatora HUMOR-a, a koji će se rabiti u računalnome morfološkom rječniku hrvatskoga standardnog jezika. Prikaz će se dati sa stajališta učenja hrvatskoga kao drugoga ili stranoga jezika te računalnoga načina obrade glagolske paradigme.

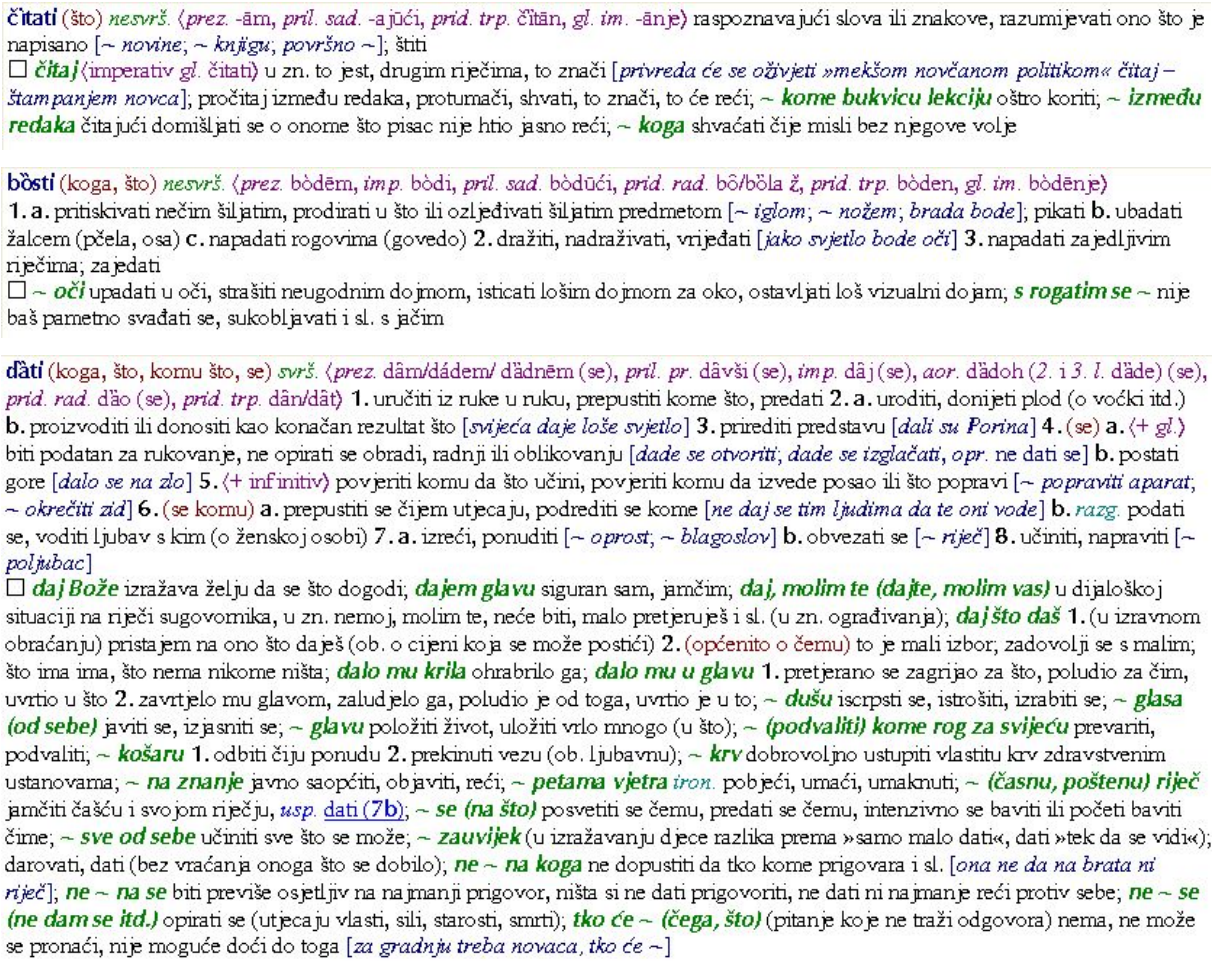
7.4.1. *Glagolska paradigma*

U Anićevu rječniku postoji 12 413 glagola. Ako se pogledaju opisi glagolske paradigme u jezičnim priručnicima za hrvatski jezik, možemo uočiti da za generiranje glagolske paradigme učenik odnosno računalo mora dobiti određene informacije da bi bilo u mogućnosti generirati glagolsku paradigmu. Slično kao i kod imenica, postupak generiranja paradigme sastoji se od nekoliko koraka. Iz razloga što je za generiranje paradigme potrebno znati glagolsku osnovu, učenik odnosno računalno nalazi se pred problemom dobivanja informacija o glagolskim osnovama. Već citirani navod iz PHG

govori o problemu generiranja osnove glagola od infinitiva te spominje i određene probleme:

> Infinitivna osnova načelno je ono što se dobije ako se odbije infinitivni sufiks. To je tako u većini slučajeva, tj. kad je ispred infinitivnoga sufiksa *–ti* neki samoglasnik. (…) Ali u ostalim slučajevima, kad imamo suglasnik (*s*) ispred sufiksa *–ti* u infinitivu ili sufiks *–ći* (…), onda treba za svaku skupinu glagola znati koja je infinitivna osnova. Infinitivna je osnova potrebna u tvorbi *aorista, radnoga pridjeva, priloga prošloga* itd.“ (Raguž 1997: 163).

Ako se povedemo za ovim citatom i pogledamo navode u Anićevu rječniku, možemo uočiti da je učeniku kojemu hrvatski nije materinski jezik, lakše pronaći glagolsku osnovu kao potrebnu informaciju u RJH nego informacije koje su mu potrebne za generiranje imenske paradigme. U prikazu 33 dane su preslike za tranzitivan glagol *čitati*, glagol *bosti* kojemu se prezentska osnova ne može dobiti ako se odbije nastavak *–ti* od infinitivne osnove te glagol *dati* koji ima i aorist i imperfekt.



čȉtati (što) *nesvrš.* ⟨*prez.* -ām, *pril. sad.* -ajūći, *prid. trp.* čȉtān, *gl. im.* -ānje⟩ raspoznavajući slova ili znakove, razumijevati ono što je napisano [~ *novine*; ~ *knjigu*; *površno* ~]; štiti
☐ ***čitaj*** ⟨imperativ *gl.* čitati⟩ u zn. to jest, drugim riječima, to znači [*privreda će se oživjeti »mekšom novčanom politikom« čitaj – štampanjem novca*], pročitaj između redaka, protumači, shvati, to znači, to će reći; ~ ***kome bukvicu lekciju*** oštro koriti; ~ ***između redaka*** čitajući domišljati se o onome što pisac nije htio jasno reći; ~ ***koga*** shvaćati čije misli bez njegove volje

bȏsti (koga, što) *nesvrš.* ⟨*prez.* bòdēm, *imp.* bòdi, *pril. sad.* bòdūći, *prid. rad.* bô/bòla ž, *prid. trp.* bòden, *gl. im.* bòdēnje⟩
1. a. pritiskivati nečim šiljatim, prodirati u što ili ozljeđivati šiljatim predmetom [~ *iglom*; ~ *nožem*; *brada bode*]; pikati **b.** ubadati žalcem (pčela, osa) **c.** napadati rogovima (govedo) **2.** dražiti, nadraživati, vrijeđati [*jako svjetlo bode oči*] **3.** napadati zajedljivim riječima; zajedati
☐ ~ ***oči*** upadati u oči, strašiti neugodnim dojmom, isticati lošim dojmom za oko, ostavljati loš vizualni dojam; ***s rogatim se*** ~ nije baš pametno svađati se, sukobljavati i sl. s jačim

dȁti (koga, što, komu što, se) *svrš.* ⟨*prez.* dâm/dádem/ dȁdnēm (se), *pril. pr.* dâvši (se), *imp.* dâj (se), *aor.* dȁdoh (*2.* i *3. l.* dȁde) (se), *prid. rad.* dȁo (se), *prid. trp.* dân/dât⟩ **1.** uručiti iz ruke u ruku, prepustiti komu što, predati **2. a.** uroditi, donijeti plod (o voćki itd.) **b.** proizvoditi ili donositi kao konačan rezultat što [*svijeća daje loše svjetlo*] **3.** prirediti predstavu [*dali su Porina*] **4.** (se) **a.** ⟨+ *gl.*⟩ biti podatan za rukovanje, ne opirati se obradi, radnji ili oblikovanju [*dade se otvoriti; dade se izglačati*, *opr.* ne dati se] **b.** postati gore [*dalo se na zlo*] **5.** ⟨+ infinitiv⟩ povjeriti komu da što učini, povjeriti komu da izvede posao ili što popravi [~ *popraviti aparat*; ~ *okrečiti zid*] **6.** (se komu) **a.** prepustiti se čijem utjecaju, podrediti se komu [*ne daj se tim ljudima da te oni vode*] **b.** *razg.* podati se, voditi ljubav s kim (o ženskoj osobi) **7. a.** izreći, ponuditi [~ *oprost*; ~ *blagoslov*] **b.** obvezati se [~ *riječ*] **8.** učiniti, napraviti [~ *poljubac*]
☐ ***daj Bože*** izražava želju da se što dogodi; ***dajem glavu*** siguran sam, jamčim; ***daj, molim te (dajte, molim vas)*** u dijaloškoj situaciji na riječi sugovornika, u zn. nemoj, molim te, neće biti, malo pretjeruješ i sl. (u zn. ograđivanja); ***daj što daš*** **1.** (u izravnom obraćanju) pristajem na ono što daješ (ob. o cijeni koja se može postići) **2.** (općenito o čemu) to je mali izbor; zadovolji se s malim; što ima ima, što nema nikome ništa; ***dalo mu krila*** ohrabrilo ga; ***dalo mu u glavu*** **1.** pretjerano se zagrijao za što, poludio za čim, uvrtio u što **2.** zavrtjelo mu glavom, zaludjelo ga, poludio je od toga, uvrtio je u to; ~ ***dušu*** iscrpsti se, istrošiti, izrabiti se; ~ ***glasa (od sebe)*** javiti se, izjasniti se; ~ ***glavu*** položiti život, uložiti vrlo mnogo (u što); ~ ***(podvaliti) komu rog za svijeću*** prevariti, podvaliti; ~ ***košaru*** **1.** odbiti čiju ponudu **2.** prekinuti vezu (ob. ljubavnu); ~ ***krv*** dobrovoljno ustupiti vlastitu krv zdravstvenim ustanovama; ~ ***na znanje*** javno saopćiti, objaviti, reći; ~ ***petama vjetra*** *iron.* pobjeći, umaći, umaknuti; ~ ***(časnu, poštenu) riječ*** jamčiti čašću i svojom riječju, *usp.* dati (7b); ~ ***se (na što)*** posvetiti se čemu, predati se čemu, intenzivno se baviti ili početi baviti čime; ~ ***sve od sebe*** učiniti sve što se može; ~ ***zauvijek*** (u izražavanju djece razlika prema »samo malo dati«, dati »tek da se vidi«); darovati, dati (bez vraćanja onoga što se dobilo); ***ne*** ~ ***na koga*** ne dopustiti da tko kome prigovara i sl. [*ona ne da na brata ni riječ*]; ***ne*** ~ ***na se*** biti previše osjetljiv na najmanji prigovor, ništa si ne dati prigovoriti, ne dati ni najmanje reći protiv sebe; ***ne*** ~ ***se (ne dam se itd.)*** opirati se (utjecaju vlasti, sili, starosti, smrti); ***tko će*** ~ ***(čega, što)*** (pitanje koje ne traži odgovora) nema, ne može se pronaći, nije moguće doći do toga [*za gradnju treba novaca, tko će* ~]

Prikaz 33: Preslika rječničkih unosa za leme *dati, bosti* i *čitati* iz RHJ

Ako se pogledaju ovi navodi, možemo zaključiti da rječnik daje nekolicinu informacija učenicima kojima hrvatski nije materinski jezik, a koje su potrebne za generiranje paradigme. Kod onih lema koje završavaju na *–ti* osnova se da neizravno zaključiti jer rječnik navodi nastavke za prvo lice jednine. Kod glagola kojima osnova, nakon što im se oduzme *–ti* završava na *–s*, naveden je cijeli oblik za prvo lice jednine (primjerice *bodem*) te se na taj način olakšava generiranje paradigme. Iz informacija koje su dostupne kod glagola *dati*, možemo vidjeti da uz lemu stoji nekoliko oblika uz pomoć kojih se generiraju različite paradigme. Iako se kod oblika *dati* ne navodi da je glagol dvovidan, navedeni su oblici i za imperfekt i za aorist.

Za generiranje glagolske paradigme učeniku su dakle dostupne neke informacije. Problem se pojavljuje u nedosljednosti prikazivanja informacija potrebnih za generiranje same paradigme. Ako samo usporedimo informacije koje su dostupne kod glagola navedenih u prikazu 34, možemo iščitati sljedeće podatke:

GLAGOL	VID	PREZ.	IMP.	PRIL.SAD./ PRIL.PR.	IMPERF./ AORIST	PRIDJ. RADNI	PRIDJ. TRPNI	GL. IM.
Čitati	nesvrš.	-am	-	-ajući	-	-	čitan	-anje
Bosti	nesvrš.	bodem	bodi	bodući	-	bo/bola ž.	boden	bodenje
Dati	svrš.	dam/dadem /dadnem (se)	daj (se)	Davši	dadoh (2. i 3. l. dade (se)	dao (se)	dan/ dat	-

Prikaz 34: Tablica sa skupnim informacijama dostupnima uz glagole *čitati, bosti* i *dati*

Kada se promotre navedene informacije, može se zaključiti da RHJ daje različite informacije uz pojedine glagole. Kod glagola *čitati* nedostaje informacija o obliku za glagolski pridjev radni, imperfekt i imperativ, glagolu *bosti* nisu pridružene informacije o aoristnome obliku, dok kod glagola *dati* nedostaje oblik za glagolsku imenicu.
Uz glagole ne postoje navodi o njihovoj pripadnosti određenoj vrsti ili razredu, tako da učenici (ili pak računalo) moraju sami doći do zaključka te potom generirati odgovarajuću paradigmu.

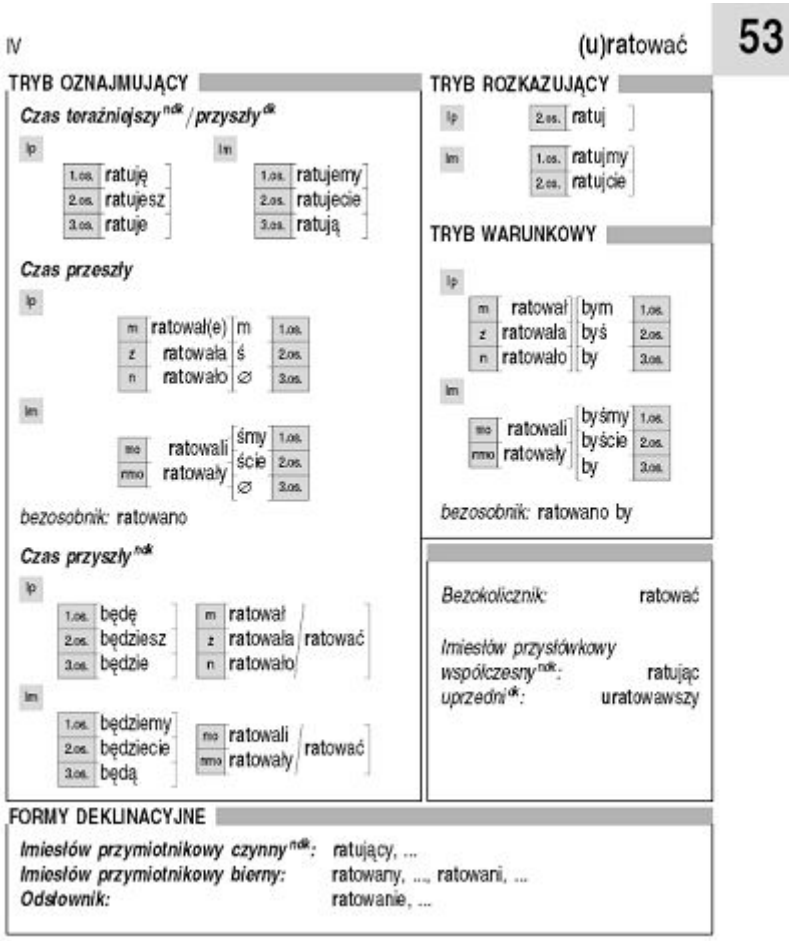
7.4.2. *Glagolska paradigma unutar morfološkoga analizatora*

Iz prethodno predočene analize možemo zaključiti da uz pomoć informacija koje su dostupne u RHJ računalo ne može generirati ispravnu paradigmu. Problem se naime pojavljuje kod lema gdje se unutar paradigme mijenja osnova. Primjerice kod glagola *voziti* u imperfektu, kod glagolskih pridjeva, te glagolske imenice pojavljuje se osnova koja završava na *–ž* (*vožah, vožaše…* i *vožen, vožena…*), dok su u Anićevu rječniku prisutni relevantni navodi samo za glagolski pridjev trpni i glagolsku imenicu.
Iz svega se ovoga da zaključiti da je potrebno osmisliti sustav koji bi uz pomoć ujednačenih parametara olakšao generiranje odgovarajuće glagolske paradigme.

Ako nadalje pogledamo neke druge slavenske jezike, primjerice ruski ili poljski, možemo uočiti da su u rječnicima uz sam glagol dane i dodatne informacije koje olakšavaju snalaženje u paradigmama te pružaju pomoć za njihovo generiranje. Prikaz 35 predočava izvadak iz Salonijeva (2001) poljskoga rječnika gdje je uz lemu navedena i glagolska skupina, čime se može doći do točne paradigme:

trapić (się) ndk t **72** ▷ **s**~
tratować (się) ndk t **53** ▷ **s**~
trawić (się) ndk t **72** ▷ **s**~
trąbić (się) ndk t **72** ▷ **za**~
trącać (się) ndk t **98** ▷ **trącić 81**
trącić *'pachnąć'* ndk it **81**
trącić (się) dk t **81** ▷ **trącać 98**
tremować (się) ndk t **53** ▷ **s**~
trenować (się) ndk t **53** ▷ **wy**~
tresować (się) ndk t **53** ▷ **wy**~
triumfować ndk it **53** ▷ **za**~
troczyć (się) ndk t **87**
troić (się) ndk t **76!**
tropić (się) ndk t **72** ▷ **wy**~
troskać się ndk it **98** ▷ **za**~ qt
troszczyć się ndk it **87** ▷ **za**~

Prikaz 35: Izvadak iz Salonijeva (2001) poljskog rječnika



Prikaz 36: Izvadak iz Salonijeva poljskoga rječnika s glagolom *trenować* i pripadajućom mu 53. glagolskom skupinom

Primjenjujući rješenja iz HUMOR-a, i u računalnome rječniku hrvatskoga jezika može se doći do sličnoga rješenja. Prvo pitanje koje se međutim postavilo jest koje sastavnice pripadaju glagolskoj paradigmi u hrvatskome standardnom jeziku.

Kod opisivanja glagolske paradigme u morfološkome parseru HUMOR-u pojavilo se pitanje obrade glagolske imenice i glagolskoga pridjeva. Pitanje se odnosilo na njihovo uključivanje pod glagolsku paradigmu ili pak pod kategoriju imenice ili pridjeva s naznakom koja ih povezuje s glagolom. Pri traženju rješenja konzultirala se sekundarna literatura, odnosno jezični priručnici. PHG naime ne spominje problematiku glagolske imenice. GHJ ne definira ovaj pojam, ali se koristi njime u nekim objašnjenjima. RHJ neke glagolske imenice svrstava pod zasebne jedinice.

Primjerice lema *čitanje* navedena je i kod glagola *čitati* kao glagolska imenica, ali i kao zasebni rječnički unos u rječniku, kao lema *čitanje* koja se povezuje s glagolom *čitati* (Prikaz 37).

čītānjle *sr* 1. ⟨*gl. im.*⟩, *v.* čitati 2. proces vezan uz praćenje teksta kao osobni ili opći interes ili interpretacija pročitanoga [*umjetničko ~e; zakonski prijedlog na drugom ~u*] 3. *kat.* odabrani ulomci iz *Svetoga pisma* koji se koriste u bogoslužju; lekcija

Prikaz 37: Glagolska imenica *čitanje* kao zaseban unos u RHJ

Za razliku od navedene glagolske imenice, glagolska imenica *unošenje* primjerice nije navedena posebno, nego samo pod paradigmu glagola *unositi* (Prikaz 38).

unòsiti (koga, što, se) *nesvrš.* ⟨*prez.* ùnosīm (se), *pril. sad.* ùnosēći (se), *gl. im.* ùnošēnje⟩, *v.* unijeti

Prikaz 38: Preslika rječničkoga unosa glagola *unositi* iz RHJ

Svrstavanje glagolske imenice u RHJ dovodi do više problema: prvi je problem određivanje broja riječi prema vrstama riječi u RHJ, što može dovesti do netočnih navoda. Iako je kod glagolske imenice *čitanje* navedena činjenica da se u 1. značenju time podrazumijeva glagolska imenica, gramatičke oznake kod same leme ne daju naznake da se ovdje radi o glagolskoj imenici, nego o imenici srednjega roda. Drugi problem koji se pojavljuje pri ovom svrstavanju glagolske imenice vezan je uz morfološki analizator. Naime, cilj je morfološkoga parsinga prepoznavanje riječi te problem svrstavanja glagolske imenice ne leži u njezinu prepoznavanju (bilo da se ona stavi pod glagolsku paradigmu ili zasebnu lemu, parser će ju prepoznati), nego se postavlja pitanje lematizacije i povezivanja riječi pri izradi prevoditeljskih alata. Naime, ako se glagolska imenica svrsta zasebno, ona se mora povezati s istoimenom kategorijom, znači glagolskom imenicom kao zasebnim rječničkim unosom. Kao primjer može se promotriti Dudenov rječnik koji je osnovica za razvijanje prevoditeljskoga alata za njemački jezik. U Dudenovu rječniku glagolska imenica *das Lesen* ne navodi se kao posebna lema, nego se smatra dijelom glagolske paradigme. Dakle, ako bismo u HUMOR-u preuzeli rječničku bazu Anićeva rječnika, ne bismo automatski mogli povezati lemu *čitanje* s njemačkom istovrijednicom, nego bismo ručno morali dodati njemačke gerunde kao zasebne leme. Ako se primijeni rješenje da se glagolska imenica stavlja pod glagolsku paradigmu, ali se pri analizi svrsta pod vrstu riječi s imenskom paradigmom, ona se pri razvijanju alata za prevođenje može lako povezati s glagolskom paradigmom koja je njezina istovrijednica te unutar nje s odgovarajućim gerundom. Time se dolazi do sljedećeg problema, a to je problem rječničke osnovice samoga analizatora, odnosno kasnije morfološkoga rječnika.

Naime, analizom RHJ nije se uočio razlog odabira pojedinih glagolskih imenica kao zasebnih rječničkih unosa u RHJ. Potragom za riječima iz prikaza 37 i 38 u HNK-u utvrđeno je da broj pojavljivanja navedenih glagolskih imenica nije zanemariv. Iako korpus nije moguće pretraživati prema lemama, nego samo oblicima riječi, utvrdilo se primjerice da se oblik *unošenje* pojavljuje 290 puta, *unošenja* 412, *unošenju* 39, a *unošenjem* 69 puta. Kod leme *čitanje* broj pojavnica prema nekim oblicima jest sljedeći: *čitanje* 1242, *čitanja* 801*, čitanju* 456*, čitanjem* 190. Prema HČR glagolska se imenica *čitanje* nalazi na 543. mjestu s apsolutnom čestoćom 27, a pojavljuje se u novinskom potkorpusu, stihovima i udžbenicima. Glagolska imenica *unošenje* nalazi se na 563. mjestu s apsolutnom čestoćom 7, a pojavljuje se u potkorpusu novina i udžbenika. Daljnjom analizom HČR utvrđeno je da su se pri korpusnoj analizi glagolske imenice svrstavale pod zasebne leme.

Rješenje koje se nametnulo kao zaključak prethodno navedenih razloga pri izradi računalnoga morfološkog analizatora jest svrstavanje glagolske imenice u dio glagolske paradigme. Njoj se pridružuje imenska paradigma, ali se pri analizi ona povezuje s glagolskom lemom.

Pri obradi glagolskoga pridjeva pojavile su se slične nedoumice. GHJ tretira glagolske pridjeve kao dijelove glagolske paradigme (2005: 41) te spominje osnovu glagolskog pridjeva radnog i glagolskog pridjeva trpnog. Kod njihove uporabe navodi da se glagoli „vrlo često ‚obezglagoljuju', tj. lišavaju i radnje i vremena. To se čini tako da se isključuju iz njih ili se pretvaraju u (glagolske) pridjeve (bilo radne bilo trpne)" (2005: 383). Znika (2005), uočavajući problem pridruživanja glagolskih pridjeva paradigmi, govori:

> Poseban se problem otvara pri analizi **glagolskih** pridjeva (radnih i trpnih). Sam naziv *glagolski pridjev radni i trpni* pokazuje dvostruku narav: s jedne strane glagolsku *(glagolski, radni, trpni)* i s druge pridjevnu *(pridjev)* koje su u disjunktivnom odnosu (odnosu ili jedno ili drugo) (Znika 2005: 431).

Na kraju svoga rada Znika (2005) dolazi do rješenja te navodi:

> Ako u glagolskom pridjevu preteže njegova glagolska sastavnica, on može biti upotrijebljen uz oblike glagola biti kojima se tvori pasiv i tada pripada glagolskoj paradigmi (…) Ako u glagolskom pridjevu (ili glagolskom prilogu) preteže njegova pridjevna sastavnica, izricanje svojstva, on može biti uvršten kao atribut i tada

> pripada imenskim riječima i ima imensku paradigmu. Tada glagolski pridjevi mogu, kao i drugi pridjevi, biti gramatičko sredstvo za izricanje određenosti imenice uz koju se uvrštavaju (Znika 2005: 438).

Tadić (1994) je pri izradi GENOBLIK-a primijenio rješenje navođenja glagolske imenice i glagolskoga pridjeva kod glagolske paradigme (Tadić 1994: 20).

Rješenje koje se primijenilo u HUMOR-u slično je rješenju za glagolsku imenicu. Glagolski se pridjev svrstava pod glagolsku paradigmu sa svojstvima pridjevske paradigme unutar računalnoga morfološkog analizatora. Pri morfološkoj analizi analizirani se oblik naznačuje kao glagolski pridjev te povezuje s odgovarajućom lemom odnosno glagolom. Izvadak iz paradigme glagola imati u obliku sustava koji koristi HUMOR dan je u prilogu 7 ovoga rada. U krajnjem lijevom stupcu dani su termovi, koji se ponavljaju u središnjem stupcu, dok desni stupac sadrži abecedne oznake kodova, kod te kraticu oblika[97].

7.4.3. *Problematika*

Jedan od problema koji se pojavljuju pri morfološkoj analizi hrvatskoga standardnog jezika, a izravno je vezan uz glagolsku paradigmu jest nemogućnost prepoznavanja složenih glagolskih vremena. Naime, razmak u HUMOR-u predstavlja granice analize. Problem je rješiv korištenjem programa HumorESK koji je obogaćen sintaktičkim znanjima (više o tome u poglavlju 9.).

Kao što je već natuknuto u prethodnim poglavljima, rješenja koja su nastala razvijanjem morfološkoga parsera HUMOR-a mogu se primijeniti na učenje i poučavanje hrvatskoga kao drugoga ili stranoga jezika. Naime, nakon definiranja sastavnica koje ulaze u sastav glagolske paradigme prišlo se kategorizaciji pojedinih glagola prema glagolskim skupinama koje se sastoje od termova te predstavljaju glagolski uzorak. Važno je pri tome napomenuti da se svaka glagolska skupina sastoji barem od jednoga različitog terma. Ukupno je utvrđeno 96 glagolskih skupina kojima se mogu pridružiti glagoli iz Anićeva rječnika.

[97] Iz razloga što cilj ovoga rada nije objašnjavanje kratica koje se rabe u HUMOR-u, neće se detaljnije objašnjavati značenja kratica koje su pripisane kodovima u prilogu 7.

U računalnome će morfološkome rječniku svakome glagolu biti pridružena glagolska kategorija. Učenici će na taj način dobiti konkretne podatke te će se moći koristiti rječnikom na sličan način kao što je prikazano u prikazima 35 i 36.

8. Računalni morfološki rječnik hrvatskoga standardnog jezika

8.1. Struktura rječnika

Uzimajući u obzir potrebe korisnika morfološkoga rječnika hrvatskoga standardnog jezika, oni se mogu grupirati na sljedeći način: korisnik, odnosno govornik kojemu hrvatski nije materinski jezik, morfološkim bi se rječnikom koristio u svrhu provjeravanja paradigme, odnosno za informacije o pojedinim oblicima riječi. Iz razloga što bi se rječnik mogao rabiti na svim stupnjevima učenja jezika, trebao bi sadržavati i istovrijednice lema na korisnikovu materinskom jeziku (ako mu materinski jezik nije hrvatski) te gramatičke oznake koje prate svaku pojedinu lemu odnosno svaki pojedini oblik (način tvorbe stema od infinitiva riječi, zatim termove za svaki pojedini nastavak, te oznake tvorbe, odnosno lemu predstavnicu određene tvorbe sa svojim obilježjima)[98]. Iz razloga što bi se rječnik rabio i za provjeravanje pravopisa, po uzoru na primjerice odrednice njemačkih rječnika, trebale bi biti naznačene i oznake za rastavljanje riječi na slogove. Međutim, iz razloga što informacija o rastavljanju riječi nema u Anićevu rječniku, uključivanje oznaka za rastavljanje riječi na slogove trebala bi biti temom nekih budućih projekata.
Ako se promotri Hrvatski morfološki leksikon, uočava se da on uz leme daje najmanju moguću količinu podataka: korijen i broj fleksijskoga/derivacijskoga uzorka. Ostale informacije (rod i informacije o tome je li riječ *singularia* ili *pluralia tantum*) daju se ako su potrebne (Tadić-Fulgosi 2003: 44).

Na temelju pretpostavljenih potreba korisnika morfološki bi rječnik sadržavao nekoliko područja:

1. Korisnik bi mogao odabrati pogled u paradigmu hrvatske riječi ili bi odabrao popis stranih riječi te odabiranjem određene leme bila bi mu ponuđena hrvatska istovrijednica čiju paradigmu želi pogledati.
2. Osim izbora leme na stranome jeziku, korisnik može birati između različitih pogleda – abecednoga poretka riječi od kraja prema početku riječi, *a tergo* ili normalnoga abecednog poretka.

[98] Problem uključivanja istovrijednica na korisnikovu materinskome jeziku zadire u pitanje leksikologije te je tema budućih istraživanja. Postoje argumenti protiv kvalificiranja ovoga rješenja rječnikom. Međutim, važno je istaknuti da ono što rječnik unutar ovoga rada može ponuditi, jest samo demo, odnosno radna inačica morfološkoga rječnika s popisom riječi na stranome jeziku. Popis je podložan daljnjemu razvijanju i dotjerivanju u smislu dobivanja potrebnih leksikoloških odrednica.

3. Nakon odabiranja određene riječi korisnik bi dobio informacije o cijeloj paradigmi koje bi se sastojale od sljedećih stavki:
    a) „sklanja se /spreže se kao X“
    b) pripada Y-toj kategoriji
    c) gramatičke odrednice (rod, broj, itd.)
    d) nepromjenjivi dio riječi (stem) tvori se tako da se od kraja rječničke osnovice odbaci Z (broj) znakova, kom se dodaju nastavci „termovi“
    e) cjelokupna paradigma.
4. Korisnik bi mogao dobiti uvid u jezičnu uporabu traženoga oblika ili tražene leme.

Za razliku od tradicionalnih morfoloških rječnika, ovaj bi rječnik hrvatskoga standardnog jezika, temeljen na računalnoj morfološkoj analizi hrvatskoga standardnog jezika, pružao drugačiji pristup paradigmama, po uzoru na slične rječnike koji postoje na svjetskome tržištu, o čemu će biti riječi u poglavlju 9.1.1. Detaljno obrazlaganje razloga izabiranja ovih mogućnosti slijedi u odlomku 8.3.

8.2. Tehnička rješenja

Uzimajući u obzir potrebe govornika opisane u prethodnome poglavlju, u ovome će odlomku biti riječi o tehničkoj izvedbi rječnika da bi se mogli ispuniti navedeni ciljevi. Pri tehničkome opisu neće biti riječi o konkretnim programima, nego o strukturi samoga rječnika i ustroju podataka.

Kao što je već navedeno, rječnik se temelji na procesu morfološke analize hrvatskoga standardnog jezika, a smatra se jednim od proizvoda procesa računalnoga morfološkog opisa hrvatskoga jezika. Iz razloga što je računalni morfološki opis hrvatskoga jezika napravljen na načelu rada s bazom podataka koja se potom prenese u oblik koji koristi morfološki parser HUMOR, baza se podataka može iskoristiti za sastavljanje morfološkoga rječnika i njegovo brzo pretraživanje nekim od pretraživačkih jezika (primjerice SQL).

Informacije koje su dostupne u pojedinim tablicama u bazi podataka gramatičke su informacije o lemama. Pri sastavljanju rječnika neke će se informacije pokazati zalihosnima te se mora izabrati ona količina informacija koja je potrebna za učenike hrvatskoga kao drugoga ili stranoga jezika. Važno je pri tome napomenuti da se ovdje radi o gramatičkim informacijama za svaku pojedinu lemu i svaki pojedini oblik, što nadopunjuje informacije koje su dostupne u Anićevu rječniku te pridonosi točnosti navoda. Ako se pogleda baza podataka, može se uočiti nekoliko tablica koje daju važne jezične informacije za sastavljanje morfološkoga rječnika. U bazi podataka postoje i druge tablice koje služe definiranju elemenata i povezivanju tablica. Za svaku pojedinu lemu daje se, osim njezina rječničkoga oblika, i *a tergo* oblik, što kasnije omogućuje abecedni poredak lema od kraja riječi, što će podrobnije biti pojašnjeno u ovome poglavlju. Osim toga, za svaku se lemu nudi broj znakova koji se moraju odbiti od kraja riječi da bi se dobio stem, odnosno nepromjenjivi dio riječi na koji se kasnije dodaju termovi. Termovi se nalaze u zasebnoj tablici s rednim brojem tipa promjene, što će kasnije biti prikazano kod prikaza paradigme, uz svaku zasebnu lemu. U jednoj su posebnoj tablici navedene informacije o vrsti paradigme, koja još jednom isključuje one oblike koji nisu zastupljeni u paradigmi. Izvadak iz radne verzije ove tablice prikazan je na prikazu 39. Važno je međutim napomenuti da se u ovim prikazima radi o radnim inačicama tablice s internim informacijama koje se neće prikazati krajnjemu korisniku.

	TID	KOD2	T	Megjegyzes
+	1	S	A	Fönév - mindent kap
+	2	S	B	Fönév - csak egyes sz mban
+	3	S	C	Fönév - csak tobbes sz mban
+	4	S	D	Fönév - mindent kap (a 22 kivetelevel)
+	5	S	E	Fönév - ragozatlan
+	6	S	F	Fönév - csak egyes sz mban (21, 31, 61 kivetelevel)
+	7	S	G	Fönév - mindent kap (a 41 kivetelevel)
+	8	V	A	Ige-ragozatlan
+	9	V	B	befejezett, nem kap: i (imperfect), n (gps -- glagolski prilog sadasni -jući)

Prikaz 39: Izvadak iz radne inačice tablice s informacijama o vrsti paradigme

Gramatičke informacije koje su dostupne za svaku lemu temelje se na informacijama iz Anićeva rječnika. Izvadak iz te tablice prikazan je u prikazu 40.

KOD1ID	KOD1	KOD	KOD2	Megjegyzes
1	Adj'	Adj'	Adj	prefix-melleknevi
2	p	Adj	Adj	melleknev
3	p indekl	Adj_Indekl	Ajd	ragozatlan meleknev
4	po	Adv	Adv	hatarozoszo (prilog)
5	indekl	Indekl	Indekl	ragozatlan szo
6	ppr	Prij	Prij	prijedlog
7	ž	Sf	S	nonemu fonev
8	m	Sm	S	himnemu fonev, nem elo
9	m1	Sm1	S	himnemu fonev, elo
10	m1 indekl	Sm1_Indekl	S	ragozatlan himnemu fonev, elo
11	m ž	Sm1f	S	himnenu es nonemu fonev-szemantikai kategoria egyarant
12	m indekl	Sm_Indekl	S	ragozatlan himnenu fonev, nem elo
13	sr	Sn	S	semlegesnemu fonev
14	sr indekl	Sn_Indekl	S	ragozatlan semlegesnemu fonev
15	pl	Splt	S	fonev, pluralia tantum
16	uzv	Uzv	Uzv	fekilatas (uzvik)
17	vezn	Vezn	Vezn	kotoszo
18	n	Vn	V	ige, befejezetlen (glagol nesvrseni)
19	s	Vs	V	ige, bejefezett (glagol svrseni)
20	zam	Zl	Z	altalanos nevmas (indefinitna zamjenica)(tko, sto)
21	zaml	ZL	Z	szemelyes nevmas (osobna zamjenica) (ja ti)
22	zamp	ZP	Z	nevmas (sebe)

Prikaz 40: Izvadak iz radne verzije tablice s gramatičkim informacijama za svaku pojedinu lemu

Tablice su povezane te se time omogućuje bolje manipuliranje podatcima. Primjer paradigme koja je povezana s tablicom, prikazan je na sljedećoj slici.

SZOID	SZO	KOD1ID	ForrasID	ZID
19	Čakovec	8	0	1
20	prostak	9	0	1
21	čovječuljak	9	0	1
22	zaselak	8	0	1
23	zadak	8	0	1
24	zadatak	8	0	1
25	nadomjestak	8	0	1
26	zalazak	8	0	1
27	vrisak	8	0	1
28	duzak	8	0	1
29	žlijeb	8	0	1
30	kalibar	8	0	1
31	kadar	8	0	1
32	izlog	8	0	1

ALAP	TID	RagozasID
1	1	47

KOD2	TERMID	TERM
S	11	g
S	21	ga
S	31	gu
S	41	g
S	51	gu
S	61	gu
S	71	gom
S	12	zi
S	22	ga
S	32	zima
S	42	ge
S	52	zi
S	62	zima
S	72	zima

Prikaz 41: Izvadak iz radne inačice tablice s lemama s mogućnosti prikaza paradigme imenice *izlog*

Kao što je u prikazu vidljivo, u radnoj se inačici baze podataka umjesto pojedinih izraza rabe numerički simboli. U morfološkome će se rječniku ovi podatci automatski zamijeniti sukladnim izrazima. Kao što je u poglavlju 4.1. već bilo objašnjeno, u računalnoj obradi hrvatskoga jezika umjesto naziva za padeže rabe se brojčane oznake. Prvi broj određuje padež, a drugi broj rod. Isto tako naveden je i uzorak paradigme (element *RagozasID*), pri čemu se navedena lema povezuje s uzrokom promjene. Ako su kod nekih oblika mogući i dvojni oblici, primjerice u N.mn. imenice *otac – oci, očevi*…, u bazi se podataka navode dvije paradigme u kojoj za svaki padež postoji samo jedan oblik (Prikaz 42).

SZOID	SZO	KOD1ID	ForrasID	ZID
11	otac	9	0	1

ALAP	TID	RagozasID
3	1	13

KOD2	TERMID	TERM
S	11	tac
S	21	ca
S	31	cu
S	41	ca
S	51	če
S	61	cu
S	71	cem
S	12	ci
S	22	taca
S	32	cima
S	42	ce
S	52	ci
S	62	cima
S	72	cima

ALAP	TID	RagozasID
3	1	14

KOD2	TERMID	TERM
S	11	tac
S	21	ca
S	31	cu
S	41	ca
S	51	če
S	61	cu
S	71	cem
S	12	čevi
S	22	čeva
S	32	čevima
S	42	čeve
S	52	čevi
S	62	čevima

Prikaz 42: Izvadak iz radne inačice tablice s lemama i njihovim paradigmama – prikaz paradigme imenice *otac*

Na temelju dobrih poveznica u bazi podataka pojavljuju se mnoge mogućnosti pri izradi morfološkoga rječnika. U sljedećem će se poglavlju predstaviti mogućnosti koje bi

morfološki rječnik trebao ponuditi da bi se osigurala učinkovitost njegova korištenja za učenje i poučavanje hrvatskoga kao drugoga ili stranoga jezika.

8.3. Mogućnosti morfološkoga rječnika

Na temelju potreba korisnika definiranih u odlomku 8. i tehničkih rješenja, u ovome odlomku slijedi detaljan opis mogućnosti koje rječnik mora ponuditi te razlozi njihova odabiranja.

Prva definirana potreba glasila je:
„Korisnik bi mogao odabrati pogled u paradigmu hrvatske riječi ili bi odabrao popis stranih riječi te odabiranjem određene leme bio bi mu ponuđen hrvatski ekvivalent čiju paradigmu želi pogledati.“

Ako se vizualni izgled rječnika izuzme iz ovoga opisa, najvažnijim se čimbenikom pri sastavljanju morfološkoga rječnika smatra njegova logičnost odnosno lakoća i preglednost njegove uporabe. Prvi problem koje se pojavljuje odnosi se na izbor jednojezične ili dvojezične inačice rječnika. Na temelju analize računalnih rječnika nekih jezika prikazane u poglavlju 9.1.1. dolazi se do zaključka da, iako se radi o morfološkome rječniku hrvatskoga standardnog jezika, uključivanje još jednoga jezika (u ovome slučaju njemačkoga) uvelike olakšava korištenje rječnika te omogućuje govornicima hrvatskoga jezika na nižim razinama jezične kompetencije njegovo lako korištenje, istodobno nudeći veliku pomoć pri učenju hrvatskoga jezika. Ako se pogledaju navodi o mogućnostima koje bi jedan računalni rječnik morao pružiti (poglavlje 9.), uočava se da je suvremena težnja pripajanje više rječnika u jednome te smanjivanje korištenja broja rječnika i dobivanje što je moguće većeg broja informacija. Time se dolazi do zaključka da bi i morfološki rječnik hrvatskoga jezika morao u ovome slučaju sadržavati njemačke leme.
Korisnik bi tada mogao izabrati između pretraživanja hrvatske leme čiju paradigmu želi pogledati, odnosno ako nije siguran u hrvatski izraz, mogao bi pogledati njemačku lemu, pri čemu bi mu se prikazala hrvatska istovrijednica.

Druga definirana potreba glasila je: „Osim izbora leme na stranome jeziku, korisnik može birati između različitih pogleda – abecednoga poretka riječi od kraja prema početku riječi, *a tergo* ili normalnoga abecednoga poretka."

Logičko i lako pretraživanje riječi važan je čimbenik koji pridonosi jednostavnosti uporabe morfološkoga rječnika. Osim već navedenoga pretraživanja prema jezicima, rječnik bi trebao ponuditi i pretraživanje prema lemama odnosno oblicima riječi u hrvatskome jeziku. Iz razloga što je rječnik namijenjen govornicima kojima hrvatski nije materinski jezik, nije potrebno postaviti mogućnost pretraživanja prema njemačkim oblicima riječi i *a tergo* funkciji. Međutim, važnim se postavlja mogućnost pretraživanja prema oblicima riječi. Pri tome se mora obaviti svojevrstan postupak lematizacije u pozadini samoga rada rječnika. Opcija pretraživanja riječi prema abecednome poretku od kraja riječi (funkcija *a tergo*) prisutna je i kod poljskoga rječnika na CD-u *Słownik gramatyczny języka polskiego*[99] (Saloni et al. 2007a), te se pokazala iznimno korisnom. Naime, polazeći od pretpostavke da slične vrste riječi imaju slične morfeme, korištenjem funkcije *a tergo* omogućuje se filtriranje lema te skupni prikaz većinom iste vrste riječi, što omogućuje lakše snalaženje u rječniku. Slično kao i kod poljskoga rječnika različite se vrste riječi mogu u popisu prikazati različitim bojama što olakšava snalaženje u popisima.
Osim pretraživanja lema, trebalo bi biti omogućeno i pretraživanje lema prema vrsti riječi, odnosno generiranje različitih popisa lema prema vrstama riječi.

Treća je potreba definirana na sljedeći način: „Nakon odabiranja određene riječi korisnik bi dobio informacije o cijeloj paradigmi koje bi se sastojale od sljedećih stavki:

a) „sklanja se /spreže se kao X",
b) pripada Y-toj kategoriji,
c) gramatičke odrednice (rod, broj, itd.),
d) neprimjenjivi dio riječi (stem) tvori se tako da se od kraja rječničke osnovice odbaci Z (broj) znakova, kom se dodaju nastavci „termovi",
e) cjelokupna paradigma.

Navedene gramatičke informacije od iznimne su važnosti za učenje stranoga jezika. Snalaženje među pojedinim paradigmama i olakšano učenje omogućeno je navođenjem

[99] Dalje u tekstu: poljski gramatički rječnik

one riječi koja ima sličnu fleksiju (*sklanja se / spreže se kao x, pripada y-toj kategoriji*). Pri tome se misli na najtipičnije predstavnike pojedinoga fleksijskog uzorka. Linkovima bi korisnik trebao moći po potrebi pogledati fleksijski uzorak leme predstavnice.
Iz razloga što je pri sastavljanju poljskoga gramatičkog rječnika uočeno da se sa stajališta učenja jezika i računalne obrade slavenskih jezika praktičnijim pokazalo korištenje računalnih kategorija, među njima i kategorije roda, jedno od rješenja koje bi se trebalo istražiti na hrvatskome jeziku korištenje je računalnih poimanja gramatičkoga roda (više o računalnome gramatičkom rodu u poglavlju 8.4.2.). U prvoj fazi izrade rječnika bilo bi poželjno koristiti se objema kategorijama – računalni gramatički rod i tradicionalni gramatički rod te provesti istraživanje o njihovoj funkcionalnosti s ispitanicima kojima hrvatski nije materinski jezik.

Kod prikaza pojedinih paradigmi potrebno je odvojiti stemove od termova lema te obvezno naznačiti broj slova koji se mora „odrezati“ od kraja riječi da bi se uvrštavanjem termova dobila točna paradigma. Termovi bi trebali biti jasno naznačeni. Ovim se postupkom primjenjuje formalizam koji je nastao unutar morfološke analize te se na ovaj način po svemu sudeći može postići formalnije objašnjavanje paradigmi u hrvatskome jeziku.

Da bi se rječnik mogao koristiti i u svrhu samoučenja odnosno samoprovjeravanja gramatičkih saznanja, može se omogućiti korištenje „učeničkoga modusa“ na dva načina:
u prvom bi se načinu omogućilo odvojeno uvrštavanje stemova u tablicu, a potom termova, što bi se pokretalo klikom miša. Drugi bi način dao priliku učeniku za samostalno upisivanje stema te njegovu provjeru, odnosno potom i samostalno upisivanje termova. Učinkovitost i korisnost navedenih načina trebalo bi prije same izrade morfološkoga rječnika provjeriti pokusnim istraživanjem.

Kao za sada posljednja potreba korisnika navodi se da bi „korisnik mogao dobiti uvid u jezičnu upotrebu traženoga oblika ili tražene leme“.
Naime, da bi korisnik dobio uvid u funkcioniranje određene leme u jezičnome korpusu, trebao bi mu se omogućiti uvid u konkordancijsku analizu korpusa. Iz razloga što je korpus vrlo velik, ta bi mu se mogućnost morala pružiti internetom. Korisnik bi time mogao upisivanjem određene riječi u tražilicu dobiti konkordanciju na nelematiziranome korpusu te time raspolagati s velikim brojem jezičnih primjera. Rad na projektu izrade rječnika kolokacija hrvatskoga jezika u svojoj je početnoj fazi te bi se s izradom morfološkoga

rječnika usporedno mogla napraviti konkordancijska analiza hrvatskoga korpusa i time korisniku pružiti dodatne korisne informacije.

8.4. Jezični problemi i rješenja koja su primijenjena pri izradi računalnoga morfološkog rječnika

Tijekom razvijanja morfološkoga parsera, kao što je već natuknuto u nekim od prethodnih poglavlja, pojavilo se nekoliko jezičnih problema koji se moraju razmotriti prije same izrade morfološkog rječnika. Iako se trudilo morfološki parser što više prilagoditi tradicionalnim rješenjima i tradicionalnoj morfološkoj analizi što se tiče gramatičkih kategorija (zadržao se primjerice tradicionalni broj padeža i oznake za pojedine padeže te rodove), ne može se zanemariti potreba pojednostavljivanja pojedinih paradigmi kod izrade morfološkoga rječnika u svrhu što lakšega i bržega učenja hrvatskoga kao drugog ili stranog jezika. Kombinacijom formalističkoga pristupa jeziku koji je nastao u okviru morfološkoga parsera i novih ideja moguće je osmisliti jedinstveni sustav koji se može primijeniti pri izradi morfološkoga rječnika, a koji bi mogao omogućiti lakše ovladavanje jezičnim sadržajima.

U ovom će se poglavlju naglasak staviti na jezičnu problematiku koja se pojavila tijekom izrade računalnoga morfološkog rječnika hrvatskoga standardnog jezika, a koji se temelji na načelima morfološke analize. Prikazat će se problematika padeža i roda sa stajališta formalističke obrade hrvatskoga jezika te će se podastrijeti moguća rješenja primjenjiva kod izrade morfološkoga rječnika. Predočena rješenja ne smatraju se konačnima, nego samo mogućima. Pokusnim će se istraživanjem unutar jednoga budućeg projekta istražiti njihova učinkovitost te učinkovitost rješenja koja su ponuđena morfološkim rječnikom.

8.4.1. *Problematika padeža*

Jedan od problema koji se pojavljuju pri obradi deklinacijske paradigme unutar računalnoga morfološkog rječnika za hrvatski standardni jezik odnosi se na uporabu svih

sedam padeža. Naime, za vrijeme izrade rječnika, pri izradi pridjevske paradigme te nakon obrade svih pridjeva koji se nalaze u Anićevu rječniku, došlo se do zaključka da na morfološkoj razini nema razlike među oblicima u dativu i akuzativu. Nakon obrade imenske paradigme došlo se do istoga zaključka te se počelo razmišljati o pojednostavljivanju paradigme te navođenju šest umjesto sedam padeža.

Naime, ako se pogleda povijesni pregled uporabe pojedinih padeža, može se zaključiti da se današnja paradigma sa sedam padeža temelji na povijesnome nasljeđu. Naime, Kolenić (2003) u svojoj knjizi govori:

> Matija Antun Reljković postavlja pitanje: „Što je casus? Was ist der Casus? Jest spadanje jednoga imena na sedam načinah…“ Dalje nastavlja s pitanjem: „Koliko u slavonskom jeziku ima casusah i kako se imenuju?“ Na ovo pitanje Reljković kao odgovor nabraja sedam padeža u jednini i množini. Premda Reljković ne ponavlja Kašićev promašaj o nejednakom broju padeža u jednini i množini, kao mnogi drugi hrvatski gramatičari koji su nasljedovali i u tome Kašića, i Reljković bez opravdanja im ablativ po uzoru na latinske gramatike, a nema lokativa (Kolenić 2003: 32).

Za razliku od Reljkovića, možemo vidjeti da Tadijanović ne označuje padeže ni imenima ni brojem, ali navodi samo šest, s padežnim pitanjima: „Tko je to? Čije je to? Komu nosiš? Koga si ubio? Koga zoveš?, Od koga ideš?“ (Kolenić 2003:120).

Ako se pogledaju morfološke karakteristike oblika u povijesti hrvatskoga jezika, možemo uočiti da je razlika između dativa i lokativa bila vidljiva samo u množini. Tako Kolenić (2003) navodi da

> Bartol Kašić u opisu sklonidbe imenica propisuje one nastavke što ih je mogao naći u djelima onodobnih pisaca. To su ponajviše nastavci čakavske i štokavske stilizacije hrvatskoga jezika. Čuva razliku u množinskim padežima dativa, lokativa i instrumentala… (Kolenić 2003: 109).

Između ostaloga on „drži da lokativ nije padež u jednini jer je po nastavcima jednak s dativom, a u množini jest, jer ima poseban nastavak“ (Kolenić 2003: 105).

Naime, Reljković nije uočio posebnost lokativa (Kolenić 2003: 120), dok Tkalčević lokativ naziva prepozicionalom (Kolenić 2003: 121).

Nadalje, analizom popisa padeža vidimo da se kod svih gramatičara koje spominje Kolenić (2003) uočavaju razlike u množini između dativa, lokativa i instrumentala (Kolenić 2003: 129).

Analizom paradigme imenica u suvremenome hrvatskome jeziku s navodima u Kolenićevoj (2003) knjizi možemo uočiti da su se danas razlike između ovih dvaju padeža na morfološkoj razini u pisanome hrvatskome standardnom jeziku izgubile. Ako se nadalje promotre rasprave jezikoslovaca o suvremenome hrvatskom jeziku, možemo vidjeti da se nekolicina bavila problemima razlike između dativa i lokativa. Raguž (1997) tako u svojoj gramatici daje sljedeću tezu:

> Iako se govori o sedam (7) padeža u hrvatskome jeziku, mnogi padeži su jednaki u pojedinim vrstama riječi (npr. u množini imenica ž. roda u DLI: *ženama*, u množini imenica m. i sr. roda u DLI: *brodovima, selima*; u množini pridjeva za sva tri roda u DLI: *lijepim(a)* itd. Međutim, u deklinaciji hrvatskoga jezika, tj. u padežnome sustavu, dva se padeža, dativ i lokativ, *ne razlikuju po nastavcima*, ni u jednini ni u množini i ni u jednom tipu deklinacije, ni u jednoj vrsti riječi (Raguž 1997: 114).

Kao što je iz ovoga citata vidljivo, Raguž (1997) prvo navodi posebne pojedine rodove da bi na kraju zaključio da su morfološki oblici jednaki u svim trima rodovima i u jednini i u množini, u svim vrstama riječi. Ovaj navod potkrepljuje sustav HUMOR gdje se obradom svih lema iz Anićeva rječnika došlo do zaključka da u suvremenom hrvatskom jeziku ne postoji razlika između dativa i akuzativa na morfološkoj razini.

Raguž (1997) međutim navodi razlike između dvaju padeža koje se očituju u uporabi navezaka i u naglasnome sustavu te govori:

> Razlike, male sasvim, ima samo po onim neobaveznim, dodatnim nastavcima, tzv. **navescima** *–e* ili *–u*, tj. *–u* je kao navezak rijedak ili nepostojeći u lokativu jednine u pridjevima i zamjenicama dok je u dativu češći. (…) Drugi, važniji i pravi razlog što se ipak govori o dva različita padeža jest razlika (u rjeđim slučajevima) u naglasku nekih imenica između dativa i lokativa jednine (Raguž 1997: 114).

Iz razloga što se sustav HUMOR koristi, kao što je već prije navedeno, za sustav morfološke analize pisanoga hrvatskoga standardnog jezika, ne uzimaju se u obzir razlike u naglasnome sustavu. Činjenica da je *–u* kao navezak rijedak ili nepostojeći u lokativu jednine, dok je u dativu češći, odnosi se na jezičnu uporabu, a ne na računalni morfološki opis hrvatskoga jezika. Jedno od osnovnih pravila jest uvođenje svih mogućih sastavnica da bi se povećala prepoznatljivost teksta. Korpusna analiza i jezična uporaba kasnije dokazuju češće ili rjeđe pojavljivanje pojedinih sastavnica.

U svome zaključivanju problematike razlike između dativa i lokativa Raguž (1997) navodi:

> Dakle, u sustavu padežnih nastavaka nema sedam različitih oblika, nego šest, ali ih je sedam kad im se pridodaju razlike po naglasku između dativa i lokativa, ali samo u malome broju riječi, i to imenica. Padežni sustav (ili sustav padežnih oblika) služi za obilježavanje odnosa među riječima u rečenici. Ali on nije sam u toj funkciji; dopunjuje ga prijedložni sustav (Raguž 1997: 115).

Primjenom ovakvoga rješenja pridonosi se lakšem usvajanju jezika jer za učenike kojima hrvatski nije materinski jezik usvajanje deklinacijske paradigme sa šest padeža predstavlja jednostavniji način usvajanja jezika. Slično sustavu njemačkoga jezika, uporaba lokativa mogla bi se pojasniti kao prijedlog + dativ. U računalnome morfološkom rječniku, da bi se ipak zadržali tradicionalni nazivi padeža u paradigmi, dativu bi se pridružila i oznaka za lokativ, kao što je naznačeno u prikazima 27 i 29.

Jedan od sljedećih problema koji su iziskivali rješenje bilo je određivanje oblika za vokativ imenica. Naime, pri obradi imenske paradigme pojavile su se određene nedoumice kod oblika za vokativ imenica koje znače neživo. Iako su neki jezičari mišljenja da vokativ tih imenica nije ni potreban, postoje dokazi o njegovoj uporabi i kod imenica za koje nije vjerojatno da bi se ikada upotrijebile u vokativu. Babić (1990) tako polemizirajući o ovoj problematici govori:

> Drugo je pitanje zovemo li mi stvari i da li nam je vokativ od imenice *pepeo* uopće potreban. Zanimljivo je da ima i lingvista koji bi na takvo pitanje odgovorili negativno. (…) Tvrdnja da nitko neće upotrijebiti koju riječ ili koji oblik veoma je osobna i praksa često opovrgava. A što je riječ manje poznata, to ćemo najprije posegnuti za priručnikom ako nam ona ustreba jer ćemo se lakše naći u nedoumici nego kod čestih i dobro poznatih riječi. Gramatike, rječnici i jezični priručnici u načelu bi trebali sadržavati odgovore na sva pitanja iz svog područja pa tako i za vokative svih imenica jer vokativ ne služi samo za dozivanje, nego i za obraćanje ili samo kao izraz osjećaja. A stvarima i govorimo i dozivamo ih jer je poosobljenje (personifikacija) česta pojava, češća nego što u prvi tren i mislimo, pogotovu u pjesničkom doživljavanju svijeta.
>
> *Zar si morala, mala željeznice,*
> *Jutros da mi odvezeš moga Javora?* (D. Tadijanović, Kad tišina spava, zlato moje)
>
> *O vjetre, brate moj, vjetre,*
> *Raskrili*
> *Golema krila svoja* (D. Tadijanović, Smrt mladića)
>
> *Izrasti, bore viti, zeleni,*
> *Iz moga srca žalosnog,*
> *Kada me u grob sahrane* (D. Tadijanović, Kad umrem)

> *Putuj! Sve! Stolico, stole, notesu, broju garderobe!* (T. Ujević, Mudre i lude djevice, 43.)
>
> *Prozore, čepe neba, odčepi mi autentični miris munja iz oblaka* (Isto.)
>
> *Gori, vatro; koltel ključaj* (W. Shakespeare, Macbeth, prev. J. Torbarina.) (Babić 1990: 139-140).

Kao što ovaj citat dokazuje, tvrdnja o nepotrebnosti vokativa za imenice koje ne znače živo nije opravdana.
Kao sljedeći problem pojavljuje se sam oblik ovih imenica, odnosno nastavci koji se rabe u vokativu jednine.

Jonke (2005) progovara o ovome problemu i navodi:

> Jasno nam je već u čemu je problem. Anti Kovačiću vokativ jednine imenice žen. roda *Laura* glasi *Lauro*, a Miroslavu Krleži *Laura*. Postavlja se pitanje koji je oblik pravilan, pa čak – u našem književnom jeziku koji ima dosta dvostrukosti – nisu li možda oba oblika pravilna (Jonke 2005: 237).

Pozivajući se za Babićevom (1990) tvrdnjom iz prethodnoga citata da bi gramatike, rječnici i jezični priručnici trebali sadržavati odgovore na sva pitanja iz svog područja, pa tako i za vokative svih imenica jer vokativ ne služi samo za dozivanje, nego i za obraćanje ili samo kao izraz osjećaja (Babić 1990: 139), smatralo se bitnim provjeriti informacije koje su dostupne o vokativu u jezičnim priručnicima koji su analizirani u poglavlju 2.3. Osim toga vokativ je oblik koji se najčešće podudara s jednim od ostalih šest oblika, tako da se njegova točnost ne može provjeriti korištenjem korpusa ili internetskih izvora.

Analizom priručnika utvrđeno je dosta nedoumica. Naime, PHG pri objašnjavanju imenske paradigme spominje da se u vokativu mogu koristiti sufiksi *–e* i *–u*. „U Vjd. alterniraju *-e* i *–u*. **Načelno**, *-e* dolazi iza nepčanih glasova, tj. iza *č, ć, dž, đ, j, lj, nj, š, ž*“ (Raguž 1997: 10). Raguž nadalje spominje izuzetke te činjenicu da „imenice koje u Njd. zvaršavaju sufiksom *–ar, -er/er, ir* (…) **načelno** u Vjd. imaju ili *–e* ili *–u*, ali je češće i običnije *–u*“ (Raguž 1997: 10)[100]. Problem se pojavljuje ako se usporede navodi u gramatici s navodima u GT, gdje kod imenica spomenutih u PHG (*gospodar, zubar)* GT navodi samo jedan nastavak u vokativu, a to je sufiks *–u*. Time učenici kojima hrvatski nije materinski jezik

[100] Isticanja moja

dolaze u nedoumicu. Naime, ako bi se učenici povodili samo za informacijama u GT, došli bi do zaključka da primjerice imenice na *–ar* mogu imati u vokativu samo nastavak *–u*. GHJ primjerice ne spominje alternacije kod sufikasa u vokativu jednine, ne definira njihovo korištenje, nego citira nekoliko primjera u kojima se mogu pročitati ili sufiks *–e* ili *–u* (Silić-Pranjković 2005:98ff). O navedenome problemu govori i Babić (1990) u svome djelu:

> U jednom se društvu povela rasprava o vokativu imenice *pepeo*. Jedni su tvrdili da je vokativ *pepele*, a drugi *pepelu*. Na kraju je jedan sudionik rekao da to nije ni važno jer je pepeo stvar koju nitko i ne zove i zato vokativa od te imenice nema i ne treba ga. Drugi se sudionik nije zadovoljio takvim zaključkom pa nas pita što mi mislimo o ovom sporu. (…) Prvo, zanimljivo je da ni u jednoj gramatici ne nalazimo vokativ te imenice, ali po gramatičkim pravilima izlazi da bi bio *pepele*, jer suglasnik *l* nije nepčani niti se imenica *pepeo* nabraja među onima koje imaju ili mogu imati nastavak *–u* s kojih drugih razloga (Babić 1990: 138).

Problem vokativa kod izrade računalnoga morfološkog analizatora i morfološkoga rječnika jest, kao što je već natuknuto, da se oblik u vokativu podudara s jednim od oblika koji se već nalaze u paradigmi, te se time korpusnom analizom ne može mnogo toga utvrditi. Ako se pogledaju jezični priručnici, uočava se da nedostaje konkretnih informacija. Naime, navodima poput *načelno mogu imati*, dopušta se korištenje obaju rješenja, ali bi se korpusnom analizom tek trebalo moći utvrditi koje je rješenje kod kojih oblika prihvatljivo, a koje nije. Pri korpusnoj bi se analizi kao jedno od rješenja mogla postaviti analiza konkordancije ili kolokacija čime bi se kao drugi token mogao zadati zarez ili uskličnik. Međutim, time bi se analiza ograničila samo na određeni broj unosa koji bi se ionako morali ručno provjeriti.

U slučaju izrade morfološkoga analizatora problem vokativa ne predstavlja preveliki problem jer će se oblici prepoznati u tekstovima te se time ne narušava postotak prepoznavanja različnica korpusa.

U slučaju izrade morfološkoga rječnika kao jedino rješenje preostaje koristiti se dostupnim izvorima i osloniti se na već postojeće sustave poput GT, jer prema riječima Stjepana Babića (1995)

> …kad se javi težnja za čuvanjem vokativa, to znači da je moramo pozdraviti. To ne znači da ja preporučujem vokativ *Hrvatsko*, želim samo reći da se ne smijemo buniti ako poimeničene riječi teže da imaju imeničku sklonidbu. Tu pojavu možemo mirne duše prepustiti praksi (Babić 1995: 170).

8.4.2. *Problematika rodova*

Kao što je već u prethodnim poglavljima naznačeno, za vrijeme razvijanja morfološkoga analizatora za hrvatski jezik te razmišljanja o prilagodbi njegovih rješenja za izradu računalnoga morfološkog rječnika za učenje i poučavanje hrvatskog kao drugoga ili stranoga jezika, naišlo se na nekoliko problema koji iziskuju dodatnu razradu. Jedan od tih problema odnosilo se i na rodove imenica.

Naime, kao što je već u poglavlju 7.2. naznačeno, ako se koriste tradicionalne morfološke kategorije, stvaraju se dodatne poteškoće pri računalnoj obradi pojedinih vrsta riječi. Pribjegavanje formalizmu i formalističkim rješenjima nudi pojednostavljivanje kako za računalnu obradu jezika, tako i mogućnosti za izradu različitih rješenja koja pridonose boljemu i lakšemu usvajanju hrvatskoga jezika. Jedno od najvažnijih problema koji su se pri tome pojavili bio je problem usvajanja rodova u hrvatskome jeziku. Važno je napomenuti da se pri tome misli na kategoriju gramatičkoga roda, a ne prirodnoga. Pregledom problematike koja se pojavila pri obradi imenica postalo je jasno da je potrebno osmisliti neko drugo rješenje. Prema uzoru na jedan od jezika kod kojega se pojavila ova problematika, naime u poljskome jeziku, pokušao se izraditi sustav prema kojemu se rodovi imenica mogu dodatno razdvojiti i detaljizirati. U ovome će se poglavlju predočiti rješenje koje se primijenilo u HUMOR-u i dati naznake za njegovo uvođenje u računalni morfološki rječnik hrvatskoga jezika.

Pitanje koje se nameće pri učenju hrvatskoga kao stranoga jezika jest „Što je rod imenica, kako se on definira i čemu služi.“ Odgovor se može potražiti u gramatikama i jezičnim priručnicima.

Ako se naime pogledaju navodi o rodovima u pojedinim izdanjima, može se vidjeti da gramatike i jezični priručnici spominju tri gramatička roda: muški, ženski i srednji. Tako PHG navodi: „Svaka imenica hrvatskoga jezika sadrži značenje **roda –** *muškoga, ženskoga* ili *srednjega*. Taj gramatički rod se ne mora uvijek podudarati s prirodnim rodom, a gramatički srednji rod, naravno, ne podudara se ni s kakvim prirodnim rodom.“ (Raguž 1995: 6). GHJ na sličan način daje informacije o rodovima imenica te daje sljedeći navod: „Imenice smo odredili kao riječi koje karakteriziraju gramatičke kategorije roda, broja i padeža. Po kategoriji roda dijele se na **imenice muškoga roda, imenice ženskoga roda** i **imenice srednjega roda“** (Silić-Pranjković 2005: 97).

Učenici kojima hrvatski nije materinski jezik, suočavaju se s poteškoćama ako žele odrediti rodove imenica, prema sljedećem navodu:

> Imenice koje znače stvari ili bića kod kojih se spol ne zna, ili nije važno da se zna, imaju rod prema svom obliku, tj. prema završetku, i to: imenice koje završavaju na suglasnik muškog su roda (osim manjeg broja takvih imenica koje su ženskog roda); imenice koje završavaju na *–a* ženskog su roda osim imenica koje znače muškarca; imenice na *–o* i na *–e* srednjeg su roda, osim vlastitih muškim imena (*Marko, Mile*) i imenice odmila (hipokoristika), koje su emocionalno obojene (*ujo, medo, braco* i sl.) (Barić et al. 1995: 101).

Naime, govornicima kojima hrvatski nije materinski jezik, a kojima nisu dostupne gotove informacije iz rječnika koje se tiču gramatičkoga roda, teško bi bilo odrediti iznimke koje se navode u gornjem citatu. Točnije rečeno, za njihovo određivanje prema gornjem navodu potrebna su semantička znanja („imenice … kod kojih nije važno da se zna [spol]“, „ženskog su roda osim imenica koje znače muškarca“, „koje su emocionalno obojene“…itd). Poteškoće se pojavljuju još i iz razloga što pojedina izdanja, poput HG, ne uvode kategorije gramatičkoga i prirodnoga roda te stoga daju složena objašnjenja poput sljedećega:

> Rod je donekle vezan i uz značenje riječi: imenice koje znače muško biće muškog su roda; imenice koje znače žensko biće ženskog su roda. Izuzeci su npr. imenice *momče* i *djevojče*, s kojima se pridjevske riječi slažu kao s imenicama srednjeg roda, i nekoliko emocionalno vrlo obojenih riječi koje znače ženska bića, a imaju oblik kao imenice muškoga roda. Pridjevske se riječi s njima slažu kao s imenicama muškoga roda. Takve su imenice npr. *djevojčurak, djevojčićak, curičak, curić, babić, curetak* (Barić et al. 1995: 101).

Analizom se ovoga navoda naznačuje da je kategorija roda kategorija koja nije izričito morfološka, što dokazuje i sljedeći citat:

> **Rod** je gramatička kategorija koja se očituje **u slaganju imenica**[101] s pridjevskim riječima, a može biti muški, ženski i srednji. (…)
> Rod je osobina imenice koja određuje da pridjevi i zamjenice koje je izbliže određuju uvijek imaju samo jedan od moguća tri oblika, tj. muški, ženski ili srednji (Barić et al. 1995: 101).

Da je određivanje roda imenica postupak koji zalazi u sintaksu, zaključak je do kojega su došli i poljski jezikoslovci te analogno tomu neizravno dali odgovor na pitanje postavljeno

[101] Isticanje moje.

na početku ovoga poglavlja, koje bi moglo glasiti: rod imenica potreban je da bismo ih pravilno slagali s ostalim riječima, da bismo dobili gramatički pravilne rečenice. Tom su prigodom poljski jezičari došli do jednoga drugog rješenja. Prema njihovim tvrdnjama, rod se imenica može lako odrediti uz pomoć sintakse, ako se imenice uvrste u jednu rečenicu. Različiti oblici pojedinih imenica, različiti nastavci ili termovi u računalnoj lingvistici ukazuju na razlike među pojedinim rodovima, što dovodi do zaključka da su potrebne dodatne potkategorije (Saloni et al. 2007a: 32-38). Što to znači za učenike kojima hrvatski nije materinski jezik i za računalno koje mora „naučiti" kako generirati pravilne rečenice?

Sa stajališta morfološke računalne obrade hrvatskoga jezika rodovi imenica naime ne predstavljaju preveliki problem jer su oni već uneseni u rječničku osnovicu programa preko Anićeva rječnika. Međutim, problem se pojavljuje u prilagodbi morfološkoga parsera nekim složenijim procesima odnosno prevoditeljskim alatima i programima za prevođenje tekstova (više o načinu rada navedenih alata u poglavlju 9.2.). Slično tomu možemo zaključiti da se učenici pri učenju hrvatskoga kao drugoga ili stranoga jezika susreću s problemom kongruencije koji nije moguće riješiti pukom uporabom kategorije rodova iz jezičnih priručnika jer se tada naime povećava mogućnost generiranja netočnih rečenica.

Primjerice, na raspolaganju učenicima (odnosno računalu u jednome širem smislu) stoje sljedeće informacije ako bismo primjerice htjeli generirati rečenice s brojevnim imenicama:

> Uz brojevne se imenice na *–ojic(a)* i *-oric(a)* u sintagmatskim odnosima rabi srednji rod množine, a uz brojevne imenice na *–oj(e)* i *–er(o)* srednji rod jednine. Usp.: *Na ulazu* ***su se našla petorica muškaraca,*** *Već* ***je stiglo petero putnika****, Pred vratima* ***se igralo troje unučadi, Ostalo*** *mu* ***je pošteđeno sedmero jaradi*** (Silić-Pranjković 2005: 145).

> Uz NAV tih brojeva obično se kaže da je riječ o Gjd. m. i sr. roda; npr. *dva učenika/kruha/djeteta/sela,* a o Nmn. imenica ž. roda; npr. *dvije djevojke/knjige.* Međutim, iz primjera u kojima uz imenicu stoji neki atribut, npr. *dva lijepa primjera, tri nova primjera, četiri plava odijela, ona dva kovčega, ova dva dobra prijatelja, moja dva sina*, očito je da takav nastavak (*-ā*) nije nikako nastavak Gjd. u deklinaciji ni određenih ni neodređenih pridjeva. Stoga to treba uzeti za zaseban oblik **dvojine** (…) stoga se u deklinaciji pridjeva mora govoriti i o dvojini (dualu), a u deklinaciji imenica ne mora – u njima je ta dvojina jednaka Gjd. za imenice m. i sr. roda, a Nmn. imenica ž. roda (Raguž 1997: 109).

Ako prema ovim navodima pokušamo generirati primjerice sintagmu s imenicom *hlače* i brojem *dva*, nailazimo na poteškoće. Naime, prema Silić-Pranjkovićevu navodu trebalo bi se raditi o srednjem rodu jednine (*dvoje hlača*), dok je iz primjera jasno da se ne radi o srednjem rodu jednine. Primjenom formalizma koji bi se temeljio na gornjem navodu generirala bi se netočna rečenica, što ukazuje i na poteškoće pri učenju hrvatskoga kao drugoga ili stranoga jezika.

Iz toga se zaključuje da je zbog problematike automatske obrade, s ciljem smanjivanja sintaktičkih pogrješaka, potrebno razviti novi sustav kojima bismo razdvojili rodove na podskupine te time pridonijeli automatskom generiranju točnih rečenica.

Primjerice, želimo li u rečenicu u sljedećem primjeru

(12) Iako [*biti*] ovdje [imenica] [odnosna zamjenica] volim, ne vidim [pokazna zamjenica] [*dva*] [imenica].

uvrstiti imenice *djeca* i *cvijeće* (obje imenice prema navodima u Anićevu rječniku imaju oznaku srednjega roda, zbirne imenice), dolazimo do sljedećih gramatički ispravnih rečenica (13) i (14).

(13) Iako **su** ovdje *djeca* **koju** volim, ne vidim **ono dvoje** *djece***.**

(14) Iako **je** ovdje *cvijeće* **koje** volim, ne vidim **one dvije [bukete]** *cvijeća***.**

Unatoč činjenici što se radi o imenicama koje u RHJ imaju istu oznaku roda (*sr.zb.*), vidimo da im je paradigma različita i da se nije moguće koristiti automatskim slaganjem jer bi se time dobili netočni navodi. Kao što se vidi iz primjera (13), zbirna imenica *djeca* ima paradigmu ženskoga roda, što se dokazuje uporabom odnosne zamjenice u ženskome rodu, ali zahtijeva kongruenciju s glagolom u množini, što je oprečno ostalim zbirnim imenicama (primjerice imenici *momčad*, koja se slaže s glagolom u jednini kao što pokazuje primjer (15)).

(15) Iako **je** ovdje *momčad* **koju** volim, ne vidim **one dvije** *momčadi*.

Ako se povedemo za gornjim navodom iz HG koji definira rod kao „gramatičku kategoriju koja se očituje u slaganju imenica“ (Barić et al. 1995: 101), zaključujemo da unutar kategorije *sr.zb.* postoje najmanje dva „podroda“, što je dokazano primjerima (13) i (14).

Nadalje, ako podrobnije pogledamo ženski rod, odnosno ponovno kategoriju zbirne imenice, te imenicu *momčad* iz primjera (15) izravno zamijenimo imenicom *janjad* koja prema RHJ ima iste oznake kao *momčad* (*ž. zb)*, dolazimo do sljedećeg primjera:

(16) *Iako **je** ovdje *janjad* **koju** volim, ne vidim **one dvije** *janjadi*.

Iako obje imenice imaju istu paradigmu (kao što je vidljivo u prikazu 43, gdje je stem od terma odvojen znakom +),

	Janjad	momčad
N	janja+d	momča+d
G	janja+di	momča+di
D	janja+di	momča+di
A	janja+d	momča+d
V	janja+di	momča+di
L	janja+di	momča+di
I	janja+di / janja+đu	momča+di / momča+đu

Prikaz 43: Računalna paradigma imenica *janjad* i *momčad*, prema GT

primjer (16) izvorni govornici ocjenjuju neprihvatljivim zbog sintagme *one dvije janjadi*, što dokazuje da se ovdje ne radi o dvama jednakim rodovima. Gramatički prihvatljiva rečenica bila bi

(17) Iako **je** ovdje *janjad* **koju** volim, ne vidim **ono dvoje** *janjadi.*

sa sintagmom *dvoje janjadi*[102]. Primjeri (16) i (17) prema tome dokazuju da se ovdje radi o dvama različitim rodovima.

Salonijevo rješenje (Saloni et al. 2007a: 32-38) poljske problematike leži u stvaranju podskupina rodova, tako da se na kraju umjesto tri roda pri računalnoj obradi jezika rabi devet (tri muška roda, jedan ženski, dva srednja i tri za pluralia tantum).
Hrvatsko rješenje naznačeno je u prikazu 44[103]:

[102] Sintagma *dvoje janjadi* pojavljuje se i u Nazorovoj pripovijetci *Otac* (Bulaja 1999) u rečenici „Kupi na silu u pastirčeta dvoje janjadi; ote nekoj ženi sa sušila nešto rublja pa joj baci izdaleka novac; ušulja se više puta noću u neke staje da pomuze krave, ostavivši na tlu zlata.“

Rečenica	Računalni rod	Primjer	Gramatički rod primjera prema RHJ
Iako **je** ovdje ___ **koji** volim, ne vidim **ona dva** ___.	m1	cvijet	m.
Iako **je** ovdje ___ **kojeg** volim, ne vidim **ona dva** ___.	m2	prijatelj	m.
Iako **je** ovdje ___ **koje** volim, ne vidim **ona dva** ___.	sr.	dijete	sr.
Iako **je** ovdje ___ **koju** volim, ne vidim **one dvije** ___.	ž.	djevojka	ž.
Iako **su** ovdje ___ **koje** volim, ne vidim **ono dvoje** ___.	pl.1.	hlače	ž. pl.tantum
Iako **su** ovdje ___ **koja** volim, ne vidim **ona dvoja**___.	pl.2	vrata	sr.pl.tantum
Iako **je** ovdje ___ **koju** volim, ne vidim **ono dvoje** ___.	zb.1	janjad	sr. zb.
Iako **su** ovdje ___ **koju** volim, ne vidim **ono dvoje** ___.	zb.2	djeca	sr. zb.
Iako **je** ovdje ___ **koje** volim, ne vidim [**one dvije bukete, ona dva para**] ___.	zb.3	cvijeće	sr. zb.

Prikaz 44: Tablica s definiranim potkategorijama gramatičkih rodova u hrvatskom jeziku

Ako usporedimo ovako dobivene kategorije rodova, koji se mogu za sada nazvati računalnim rodovima, vidimo da se poneki gramatički rodovi razlikuju međusobno. Primjerice, iako kod potkategorija zb.1, zb.2 i zb.3 RHJ navodi samo kategoriju srednji rod, zbirna imenica, možemo vidjeti da se prema sintaktičkim pravilima radi o trima različitim rodovima. Ako nadalje usporedimo imenice iz RHJ te njihove gramatičke oznake rodova s novonastalim potkategorijama, dolazimo do sljedećih zanimljivih opažanja. Naime, imenicu *momčad* RHJ obilježava kao *ž. zb.* Prema GT ta imenica nema množinu, odnosno može se pojmiti u drugome slučaju kao množina imenice *momče*, s paradigmom koja je istovjetna jednini u slučaju kada se prema GT radi o zbirnoj imenici bez množine. Analogno tomu mogli bismo zaključiti da ne postoji sintagma *dvije momčadi*. Međutim, provjerom navoda u HNK zaključuje se da se imenica momčad u sintagmi *dvjema momčadima* pojavljuje u tri navrata. Time se zaključuje da je rod imenice ženski (računalni rod *ž*) te da se u ovome slučaju u gramatičkome smislu radi o imenici koja ima svoju množinu.

Sljedeće opažanje odnosi se na imenicu *tata*. Iako bi se prema termovima za imenicu *tata* moglo reći da je istovjetna imenici *mama*, koja je ženskoga roda, jer imaju istu paradigmu i

[103] Iz razloga što hrvatsko rješenje nije bilo potrebno pri izradi morfološkoga analizatora, predlagano rješenje u prikazu 44 ne smatra se konačnim, nego podložnim promjenama uslijed izrade računalnoga morfološkog rječnika hrvatskoga standardnog jezika odnosno daljnje prilagodbe programa prevoditeljskim alatima.

u jednini i u množini, kongruencijom se dokazalo da se radi o imenici muškoga roda koja pripada potkategoriji m2 iz prethodne tablice. Ovo još jednom potvrđuje činjenicu da je rod kategorija koje se definira na temelju sintakse. Analogno tomu, kao sljedeće opažanje pojavljuje se imenica *Hrvatina*. Naime, prema navodima iz RJH radi se o imenici muškoga roda. Ako se pak pogleda jezična uporaba, vidi se da se riječ ponajviše upotrebljava u sintagmi koja upućuje na to da se radi o ženskome rodu (*ove Hrvatine, jedna Hrvatina*[104]) potkategorije ž.

Sljedeća imenica koja je podvrgnuta testiranju jest imenica *prsa*. Naime, u RHJ uz ovu imenicu stoji oznaka *ž. pl. tantum*. Uvrštavanjem imenice u navedeni test zaključuje se da je imenica istoga roda kao *vrata*, koju RHJ tretira kao *sr. pl. tantum*. Naime, testiranjem internetskih stranica utvrđeno je da se upotrebljava sintagma *dvoja prsa*, ponajviše u obliku *dvoja pileća prsa.* Ako se nadalje testira i imenica *leđa*, utvrđuje se da se radi također o potkategoriji p2.

Određivanje ovakvog načina bilježenja rodova pokazalo se korisnim i u slučajevima kada se radi o imenicama koje se navodno ne dekliniraju, primjerice vlastitom imenu Ines. Naznakom računalnoga roda učenik će biti u stanju pravilno ju složiti s ostalim rečeničnim sastavnicama. Kao drugi sličan primjer koji se može navesti jesu imenice koje se podudaraju svojim oblicima, primjerice *vodič* (u značenju osobe) i *vodič* (u značenju knjige koja sadržava upute i podatke). Ako pogledamo prethodnu tablicu, možemo vidjeti da prema uvrštavanju u testnu rečenicu ove dvije riječi pripadaju dvjema kategorijama.

Jedan od problema koji se naime pojavljuje ako se primijeni ovo rješenje, odnosi se na naznaku roda u morfološkome rječniku i priručnicima za učenje hrvatskoga kao drugoga ili stranoga jezika. Ako se navedeni test pokaže točnim, može se tvrditi da u hrvatskome jeziku postoji više od tri roda, što povlači za sobom cjelokupnu razradu novoga pristupa učenju i poučavanju hrvatskoga jezika. Naime, ako bismo saželi dosadašnja saznanja, mogli bismo zaključiti da bi se u svrhu poboljšanja korištenja jezičnih priručnika i rječnika natuknice u rječnicima morale sastojati od dviju informacija. Prva bi sadržavala term i fleksijsku skupinu, na temelju čega bi učenici (ili računalo) mogli generirati odgovarajuću paradigmu, kao što je primjerice prikazano u poglavlju 7.3. Druga bi sadržavala primjerice računalni rod (primjerice kod imenica) na temelju koga bi učenik mogao generirati

[104] Iz razloga što se imenica *Hrvatina* ne pojavljuje u HNK, pretražile su se internetske stranice, gdje se sintagma koja upućuje na ženski rod pojavljuje u više navrata od sintagmi koje upućuju na muški rod imenice.

pravilnu paradigmu. Ovim dvjema informacijama dao bi se odgovarajući temelj za uspješno svladavanje fleksijske paradigme i generiranje gramatički što ispravnijih rečenica.

9. Ostali oblici primjene morfološke analize

Ako se pogledaju dosadašnji radovi i dosadašnja dostignuća na području računalne lingvistike, može se uočiti da, osim izrade morfološkoga rječnika hrvatskoga standardnog jezika, postoje mnoge industrijske aplikacije računalne obrade sustava nekoga jezika, konkretno u ovome slučaju procesa morfološke analize, odnosno sinteze. U ovome se dijelu rada nudi pregled nekih dosadašnjih dostignuća i istraživanja, odnosno prikaz jezičnih alata i programa koji se u nekoj svojoj fazi koriste postupkom morfološke analize. Iz razloga što je tema ovoga rada analiza prema načelu rada morfološkoga parsera HUMOR-a, usredotočit će se na jezične alate i programe koji se koriste navedenim analizatorom odnosno njegovom verzijom pojačanom sintaktičkim znanjem, HumorESK[105]. Osim toga dat će se i teorijski, jezikoslovni pregled razvijanja nekih od navedenih programa ili jezičnih alata za hrvatski jezik. Cilj je ovoga poglavlja prikazati načine korištenja postignuća u računalnoj lingvistici te mogućnosti iskorištavanja rezultata istraživanja ne samo na području translatologije, nego i na polju razvijanja nastavnih materijala za učenje i poučavanje kako hrvatskoga, tako i stranih jezika.

Kao što je već poznato, računalna su se dostignuća već u ranoj fazi počela razvijati i unutar jezikoslovlja, odnosno humanističkih znanosti. Prema riječima Prószékyja (2003) jezične tehnologije ne odnose se na činjenicu kako računalo pomaže humanističkim znanostima, odnosno kako računalno pomaže rad jezikoslovca, nego kako se lingvistički rezultati mogu primijeniti i biti dostupni računalu. Drugim riječima, jezične tehnologije ne znače računalo u jezikoslovlju, nego jezikoslovlje u računalu (Prószéky 2003: 5). Međutim, prije samoga opisa primjene računalnih dostignuća važno je definirati pojmove o kojima će biti riječi u sljedećim odlomcima, a to su *alat* i *softver (programi)*. Prema definicijama, *jezični* ili *računalni alat* računalni je program pomoću koga se obavlja složeni zadatak ili više njih. *Softverom* se u ovome smislu podrazumijeva paket programa ili engl. *software package* koji sadrži više alata spojenih u jedan sustav[106]. U ovome će se radu umjesto pojma *softver* rabiti hrvatska inačica koju predlaže Mihaljević (1993), *programi* (vidi Stojaković 2004: 86).

[105] U ovome će se poglavlju dati pregled samo onih alata i programa koji se koriste navedenim parserima te za koje postoji realna mogućnost da se prilagode i hrvatskom jeziku. Za više informacija o svim ostalim proizvodima koji se koriste morfološkim parserom HUMOR na internetskoj stranici www.morphologic.hu

[106] http://wiki.answers.com/Q/What_is_the_difference_between_software_tools_and_software_packages, 21. sprnja 2008.

Ako se pak nadalje promotre računalno-lingvistička dostignuća unutar jezikoslovlja, može se uočiti da rezultati nailaze na široku primjenu na različitim poljima.

Zasigurno se kao prvi način uporabe pretpostavlja korištenje morfološke analize u svrhu razvijanja programa za rastavljanje riječi na slogove (više o programima za rastavljanje riječi na slogove u Prószéky 1999a).

Osim toga, postupkom se analize uvelike koristi i kod programa za sažimanje, odnosno alata za izvlačenje riječi iz nekoga teksta. Njime se naime omogućuje generiranje popisa s riječima iz teksta koji je podložan daljnjoj obradi u smislu davanja prijevoda lema ili pak statističke analize poput analize konkordancije ili kolokacija. U hrvatsko-njemačkome okviru time bi se omogućio dodatan napredak što se tiče razvoja kvalitetnijih i pouzdanijih programa koji bi bili od velike koristi kako hrvatskim, tako i njemačkim govornicima.

Kao što je već u poglavlju 2.1.3. pojašnjeno, postupkom morfološke analize omogućuje se bolja uporaba programa za statističku analizu teksta koja se tiče traženja kolokacija ili konkordancija riječi. Analizom lematiziranoga teksta dobivaju se pouzdani rezultati koje je moguće iskorisiti u druge svrhe, poput primjerice sastavljanja rječnika kolokacija za učenike kojima hrvatski nije materinski jezik. Isto bi se tako omogućilo i sastavljanje posebnih rječnika koji se temelje na čestotnosti pojavnica u specijaliziranim korpusima ili potkorpusima.

Procesom morfološke analize, odnosno korištenjem morfološkoga analizatora na opsežnome korpusu ili pak korpusu suvremenoga jezika, stvara se osnova za dodatna istraživanja traženja novih riječi koja su nadalje podložna daljnjoj obradi. Naime, ako pretpostavimo da rječnička osnovica morfološkoga analizatora sadrži i prepoznaje samo riječi koje su u sastavu Rječnika hrvatskoga jezika, pretraživanjem korpusa neprepoznatljivima i neobrađenima ostaju one riječi koje nisu u sastavu rječnika. Ti su oblici nadalje podložni daljnjoj ručnoj obradi u smislu čišćenja od suvišnih sastavnica ili pak određivanja kojem jeziku ili jezičnoj varijanti pripadaju. Za one riječi za koje se utvrdi da su sastavnice hrvatskoga jezika, ovim se načinom lako može utvrditi radi li se o novokovanicama ili pak riječima koje su možda ostatci srpsko-hrvatskoga jezičnog naslijeđa, ali su još uvijek u jezičnoj uporabi. Time se zaključuje da se morfološkom analizom mogu pronaći i odabrati nove riječi te napraviti odgovarajuća istraživanja.

Morfološka se analiza može koristiti i na način da ju se ugradi u različite tražilice. Naime, pri traženju određene riječi, kolokacije ili izraza i upisivanjem riječi u tražilicu pretražuju se tekstovi samo s onim oblikom koji je zadan. Uporabom postupka morfološke analize pretraživač bi nakon eventualnoga analiziranja riječi mogao generirati cjelokupnu

paradigmu te u traženje inkorporirati sve oblike koji su sastavni dio paradigme. Primjerice, ako bismo u tražilicu upisali lemu *jabuka*, pretraživač bi nam osim rezultata s oblikom *jabuka* ponudio i rezultate s oblicima *jabuke, jabuci, jabuku, jabuko, jabukom* i *jabukama*. Slično tomu, ako bismo upisali određeni oblik riječi u tražilicu te ga označili znakom koji bi podrazumijevao da se radi o jednome obliku i time dali znak pretraživaču da analizira oblik, generira paradigmu i traži sve oblike koji se nalaze u paradigmi te nam ponudi odgovarajuće rezultate, mogli bismo pretraživati veliki korpus te na brži način dobiti rezultate koji nas zanimaju.

Ipak, među poljima u kojima se koristi morfološka analiza prednjači translatologija. Naime, marljivo se razvijaju različiti alati i programi koji se na različite načine rabe pri prevođenju, odnosno kao pomagala prevoditeljima, o čemu će više riječi biti u ovome poglavlju. Osim translatologije računalne tehnike igraju veliku ulogu i na polju leksikologije i leksikografije te u razvijanju alata koji se koriste pri učenju i poučavanju jezika. Na temelju razmišljanja u Prószékyjevim radovima (1997c i 1997d) zaključujemo da računalni programi i alati polako počinju zamjenjivati klasične oblike rječnika, leksikona i tezaurusa te predstavljati moderniji oblik sa svim odrednicama koje imaju njihove „papirnate“ inačice, ali u isto vrijeme nudeći mnogo više mogućnosti nego što je to tehnički izvedivo u klasičnome papirnatom obliku. Oblici njihove primjene prelaze granice prevođenja i leksikologije te ulaze čak i u područje učenja i poučavanja jezika, odnosno već u prethodnim poglavljima spomenute discipline CALL (Computer-aided Language Learning) te DDL (Data-Driven Language Learning). U sljedećim će se odlomcima podrobnije objasniti navedena dostignuća te dati pregled primjene jezičnih tehnologija na ovim dvama poljima.

9.1. Jezični alati za učenje i poučavanje jezika

U današnje se vrijeme na tržištu mogu pronaći različiti programi koji se mogu koristiti u svrhu učenja i poučavanja jezika. Definicija same discipline CALL-a može se sažeti u sljedećih nekoliko redaka:

> U posljednjim godinama sve više alata pokušava koristiti *tehnike profesionalnoga jezičnoga inženjeringa,* posebice morfosintaktičke obrade i korpusne analize, u

> tehnologiji *učenja jezika uz pomoć računala*. Zbog toga se u današnje vrijeme izraz jezična tehnologija (LT), odnosno obrada prirodnih jezika (NLP) + industrijska tehnologija rabi češće od izraza NLP. Drugim riječima, postaje zavaravajuće nazivati dobropoznatu tehnologiju CALL-a tehnologijom učenja jezika uz pomoć računala, kada se zapravo radi o poučavanju jezika uz pomoć računala (*computer-aided language teaching* ili *CALT*). CALT, dakle, nije ništa drugo nego tradicionalno poučavanje jezika (TLT) koje podupiru programski alati.[107] (Prószéky 1997b: 53).

Ako se podrobnije prouče programski alati koji su danas prisutni na tržištu, zaključuje se da se oni daju podijeliti na nekoliko kategorija:

1. programi za strojno prevođenje,
2. rječnici,
3. alati za provjeru pravopisa, tzv. *proofing tools,*
4. softver za učenje jezika.

U skupinu alata koji se danas rabe unutar discipline CALL ili CALT ubrajaju se prije svega alati u obliku jednojezičnih ili višejezičnih računalnih rječnika koji se u nekoj od svojih faza koriste morfološkom analizom ili sintezom, lematizacijom ili disambiguatorima. Osim navedenih alata, kao važnim se sredstvima, ne samo pri učenju i poučavanju jezika, smatraju i različiti alati za provjeravanje pravopisa i gramatike, rastavljanja riječi, a koji se ponovno temelje na navedenim postupcima.

Među jezičnim alatima kojima je osnovica morfološka analiza odnosno sinteza ima mnogo uspješnih alata koji se u nekoj svojoj fazi koriste morfološkim parserom HUMOR-om (Prószéky 1994). Važno je prije opisa samih alata napomenuti činjenicu da se predstavljanje navedenih dostignuća ne smije shvatiti kao reklamna, nego kao informativna svrha čiji je cilj davanje informacija o različitim mogućnostima korištenja postupka morfološkoga parsinga. Nipošto se ne smatra da je HUMOR odnosno da su alati koji potiču od proizvođača, tvrtke MorphoLogic, isključivi, nego je predstavljanje navedenih alata izabrano iz razloga što su oni dobitnici trinaest međunarodnih nagrada, od kojih se

[107] Prijevod MA. Izvornik: „In the recent years more and more software tools aim to apply professional language engineering techniques, especially morpho-syntactic processing and corpus analysis, to technology for computer-assisted language learning. This is why nowadays the term language technology (LT), that is, natural language processing (NLP) + industrial technology, rather than NLP, is generally used. In other words: it seems a misleading terminology to call the well-known discipline computer-aided language learning (CALL), because it is, in fact, computer-aided language teaching (CALT). CALT is, therefore, nothing else but traditional language teaching (TLT) supported by software tools."

izdvaja nagrada za najinovativniji računalni alat (IST-Prize koji dodjeljuje Euro-CASE)[108]. Osim toga, postoje realne mogućnosti pokretanja projekta prilagođavanja ovih jezičnih alata i hrvatskomu jezičnom sustavu.

9.1.1. *Jednojezični i dvojezični računalni rječnici*

Kada je riječ o jednojezičnim ili dvojezičnim računalnim rječnicima, postoji nekoliko načela pri njihovoj izradi. Većina računalnih rječnika dostupnih preko CD-roma kao medija (poput primjerice Anićeva *Rječnika hrvatskoga jezika* na CD-romu) nudi samo one mogućnosti koje postoje u papirnatome izdanju, odnosno pretraživanje samo prema lemama te jednodimenzionalni prikaz objašnjenja, preslika tiskanoga izdanja, što je detaljno predočeno u poglavlju 2.3.5. Prednosti koje takav oblik računalnoga rječnika nudi, očituju se samo u brzini pretraživanja lema, odnosno uštedi korisnikova vremena. Rječnik ne nudi nikakve dodatne mogućnosti, primjerice osvježivanje baze podataka, koje nudi primjerice njemački Dudenov računalni rječnik njemačkoga jezika *Großwörterbuch der deutschen Sprache*. Prószéky pri razvijanju računalnih rječnika potiče inkorporiranje suvremenih dostignuća na području računalne tehnologije te se s tim u vezi zalaže za sljedeće načelo:

> Na temelju toga, u budućnosti moramo kombinirati obje gore navedene potrebe: načiniti leksikološke izvore računalno dostupnima te pokazivanjem strategije kako ih razviti pokušavamo unijeti promjene u razvojne strategije elektronskih rječnika. Današnja se tehnologija može – i mora – koristiti dinamičnim sadržajima, poput morfo-sintaktičke analize, lematizacije, provjere pravopisa itd. S druge strane, rječnici nikada ne mogu biti potpuni bilo u kojem smislu, tako da moramo omogućiti pristup usporednim multirječnicima. To znači da se jedan rječnički ulaz mora koristiti neograničenim brojem leksičkih izvora koji su dostupni prevoditelju[109] (Prószéky 1999c: 221).

[108] http://www.morphologic.hu/index.php?option=com_content&task=view&id=616&Itemid=238#9, 21. srpnja 2008.

[109] Prijevod MA. Izvornik: „Consequently, in the near future we have to combine the two above needs: making existing lexical resources computationally accessible and showing the strategy how to develop we try to argue for changes in development strategies of electronic translation dictionaries. Today's language technology can – and must – use dynamic actions, like morpho-syntactic analysis, lemmatization, spell checking, and so on. On the other hand, dictionaries can never be full in any sense, therefore we have to make parallel multidictionary access possible. It means that a single dictionary look-up should use an unlimited number of lexical resources that are available for the translator."

Interpretacijom prethodno navedenoga citata možemo zaključiti da razvoj rječnika prati potrebe korisnika kojima je cilj u što kraćem vremenu doći do što više informacija o nekom leksemu. Informacije prvenstveno prate kognitivne procese koji se pojavljuju pri traženju značenja nekoga leksema, što znači da i računalni rječnici moraju odgovoriti potrebama korisnika te time inkorporirati nekoliko postupaka:

1. morfološku analizu, odnosno identifikaciju zadane riječi,
2. lematizaciju (odnosno povezivanje oblika riječi s lemom, odnosno natuknicom u rječniku),
3. identifikaciju vrste riječi (POS-disambiguator),
4. semantički disambiguator (da bi se odredilo značenje),
5. etimološki rječnik,
6. tezaurus (za određivanje sinonima),
7. sintaktički parser (za određivanje funkcije zadane riječi u rečenici),
8. hyphenator (za određivanje načina rastavljanja riječi),
9. izgovor,
10. slikovni rječnik,
11. provjeravanje pravopisa, gramatike i objašnjenje pravopisnih pravila,
12. korištenje riječi u zadanome korpusu,
13. dodavanje ostalih rječnika ili enciklopedija i njihovo slobodno integriranje u sastav rječnika (MorphoLogic je razvio slobodno dostupan sustav MobiGloss, koji je u mogućnosti pretraživati traženu riječ na internetu ili pak u MorphoLogicovoj bazi podataka na poslužitelju (Prószéky 1997b).

Sve su od ovih navedenih funkcija sadržane u MorphoLogicovu rječniku MobiDicu.
MobiDic je komercijalno ime za *multi-dictionary environment*, odnosno računalni rječnik koji se može koristiti u oba smjera. Polazišni jezik može biti i ciljni jezik. Prvenstveno je razvijen za korisnike koji se koriste internetom ili intranetom (Prószéky 1999: 221). Do sada su razvijene verzije MobiDica za engleski, njemački i mađarski jezik, ali eksperimentalne verzije uključuju i španjolski i poljski (Prószéky-Kis 2002: 284). Za razliku od drugih računalnih rječnika, MobiDic sadrži kompilaciju više rječnika i jezičnih alata. To znači da se jedna lema ili jedan leksem ne pretražuje u samo jednome, nego u više rječnika, te se korisniku analogno tomu nude različita rješenja. Primjerice, korisnik može birati između slikovnoga rječnika sa zadanim jezikom kao polazišnim ili ciljnim

jezikom ili nekoga drugoga rječnika (Prószéky-Kis 2002: 284). Važno je napomenuti da sustav ne nudi samo pretraživanje prema lemama, nego različitim oblicima riječi, bez obzira pripada li riječ jednoj skupini, frazi ili ne, jer (Prószéky-Kis 2002b: 284) se na rječničkoj podrazini događaju procesi lematiziranja, morfološkoga parsinga, odnosno diasmbiguacije (Prószéky 1999: 222). Tehničke karakteristike isto tako izdvajaju MobiDic od ostalih računalnih rječnika prisutnih na tržištu.

MobiDic je naime rječnik sa strukturom klijenta-servera, što znači da je pohranjen na jednome serveru prevoditeljske skupine ili tvrtke. Računala korisnika samo se koriste navedenim programom odnosno mogu pretraživati i unositi nove rezultate, a dostupni su im svi rječnici koji se nalaze na središnjem serveru. MobiDic tako podržava OS Windows NT, Unix (Linux, Solaris) Windows 95, 98, NT i XP, a korisnici mu mogu pristupiti preko TCP/IP protokola (Prószéky-Kis 1999b).

Razvijanjem jezičnih alata za hrvatski standardni jezik omogućila bi se i izrada dvojezične verzije MobiDic-a, što bi pridonijelo i razvoju hrvatsko-mađarske leksikografije.

Kao drugi alat koji je višestruko nagrađivan, a koji u novije vrijeme u svakome pogledu odgovara potrebama korisnika jest MobiMouse. Jedan od glavnih ciljeva pokretanja projekta morfološke analize hrvatskoga standardnog jezika bilo je upravo razvijanje jednoga takvoga jezičnog alata koji se temelji na HUMOR-u i za hrvatski jezik. MobiMouse kao proizvod tvrtke MorphoLogic za razliku od sličnih proizvoda na tržištu raspolaže s mnogim prednostima. Ovaj se jezični alat naime definira kao *context-sensitive instant comprehension tool*, odnosno program koji ujedinjuje nekoliko postupaka – morfološku analizu, lematizaciju i disambiguaciju. Jednom pokrenut, alat nenametljivo radi u računalnoj pozadini te postavljanjem kursora bilo na koju riječ koja se pojavljuje na korisnikovu zaslonu nudi u zasebnome balončiću odgovarajući prijevod kako leksema, tako i skupine riječi, ako je riječ dio fraze. Kategorizacija *context-sensitive* pripala mu je iz razloga što ne nudi sve moguće prijevode neke riječi, nego na temelju različitih postupaka analizira sam kontekst u kojem se određeni leksem pojavljuje te tako nudi samo odgovarajući prijevod. Naime, sustav nadilazi sustav računalnih rječnika jer rječnički ulaz dovodi do točke prijevoda. Ipak, definira se manje od sustava za prevođenje jer ne postoji sintaktička obrada polazišnoga teksta, nego samo niz rječničkih provjera (Prószéky-Kis 2002: 281). Ako se ovaj računalni alat usporedi sa sličnim pop-up rječnicima poput Babylona, WordPointa, CleverLearna, iFingera, Langenscheidtova Pop-up Dictionaryja i Techocrafova RoboWorda sa stajališta funkcionalnosti i jezične točnosti, može se vidjeti

da je MobiMouse u prednosti zbog veće količine ponuđenih mogućnosti (Prószéky-Kis 2002: 285). Osim morfološkoga parsinga koji se događa na načelu HUMOR-a, alat radi u tri faze. Cilj je prve faze jezični *stemming* ili postupak povezivanja oblika riječi s rječničkim oblikom leme, zatim provjeravanje pravopisa, dok se u trećoj fazi provodi parsing na temelju konteksta da bi se identificirale eventualne skupine riječi. Prijevodi potječu iz različitih izvora jer je alat, slično MobiDicu, u kratkome roku u mogućnosti povezati nekoliko različitih rječnika (vid Prószéky-Kis 2002).

Tehničke prednosti MobiMousea uključuju i brzo provođenje zadataka. Brzina modula za prepoznavanje teksta jest 1000 znakova/s, stemming se događa za 0,0002s/oblik riječi, rezultati se prikazuju u vremenu od 0,02s (Prószéky-Kis 2002: 290). U svrhu prepoznavanja teksta MorphoLogic je razvio identifikator jezika pod nazivom LangWitch. Međutim, pri prepoznavanju jezika, prije no što počnu spomenuti postupci, rabi se jednostavnija metoda, a to je metoda pretraživanja rječnika ovog jezika. Ako je riječ prisutna u jednom od rječnika, ona je prepoznata, obrađuje se i naposljetku prevodi (Prószéky-Kis 2002: 284-285). Činjenice da su kod MobiMousea jezici reverzibilni, da je sam program neupadljiv i ne ometa rad drugih aplikacija na računalu, daju prednosti ovom alatu pred ostalim sličnim pomagalima (o ostalim tehničkim karakteristikama alata više u Prószéky-Kis 2002).

Jedan od rječnika za učenje slavenskih jezika koji je dostupan na CD-romu gramatički je rječnik poljskoga jezika (Saloni et al. 2007b). Poljski je jezik sa stajališta računalne obrade slavenskih jezika zanimljiv iz razloga što se pri njegovu računalnom opisu moraju rješavati problemi koji su slični hrvatskome jeziku, a za razliku od nekih slavenskih jezika poljski se jezik ne koristi ćiriličnim pismom.

CD-rom izdanje *Słownik gramatyczny języka polskiego*-a, prema riječima autora, temelji se na morfološkoj analizi, a sadrži riječi koje se nalaze u rječnicima i u korpusu (Saloni et al. 2007a: 155). Programska mu je osnovica sustav GNU/Linux i Ubuntu. Za razliku od Anićeva rječnika, poljski gramatički rječnik može se koristiti interaktivno te nudi poredak lema prema abecedi od početka i od kraja riječi, što je korisno ako se uzme u obzir hipoteza da slični sufiksi upućuju i na slične ili pak istu vrstu riječi (primjerice pridjevski nastavci u hrvatskome jeziku).

Ako se pogleda popis lema s lijeve strane rječnika, vidimo da su različite vrste riječi odvojene različitim bojama, koje se pojavljuju i kod opisa pojedinih paradigmi. Kod svake pojedine natuknice poljski gramatički rječnik nudi informacije o vrsti riječi, broju

paradigme te rodu riječi. Zanimljivo je uočiti da su kod razvijanja poljskoga gramatičkog rječnika njegovi tvorci došli do novih saznanja što se tiče rodova pojedinih riječi, što se može primijeniti i na hrvatski jezik (više o tome bilo je riječi u poglavlju 8.4.2.).

Kod samoga prikaza paradigmi moguć je prikaz prema oblicima riječi. Rječnik nudi mogućnost prikaza svih oblika unutar određene paradigme ili prikaz samo različitih oblika u paradigmi, što omogućuje brže snalaženje u rječniku te brže pronalaženje potrebne informacije.

Osim upisivanja tražene riječi, rječnik nudi i poseban prikaz lema prema različitim kriterijima (primjerice vrstama riječi) te istodobno prikazuje i broj lema koje se pojavljuju unutar svake pojedine vrste riječi.

Mogućnosti koje rječnik pruža mogu uvelike pridonijeti uspješnijem svladavanju poljskoga jezika, a rješenja i ideje mogu se uspješno primijeniti i na sastavljanje morfološkoga rječnika hrvatskoga standardnog jezika.

9.1.2. *Jezični alati za učenje mađarskoga jezika*

Kao što je više puta spomenuto, jednim od ciljeva ovoga rada postavila se izrada i opis morfološkoga rječnika hrvatskoga standardnog jezika koji se temelji na načelu morfološke analize. Za razliku od slučaja hrvatskoga jezika gdje nema mnogo sličnih alata za učenje jezika, a koji bi se mogli uspješno primjenjivati u nastavi hrvatskoga kao drugoga i stranoga jezika, u Republici Mađarskoj na tržištu su se našli programi koji se temelje na odrednicama morfološke analize i sinteze mađarskoga jezika. Među takve programe ubrajaju se i oni koji se temelje na morfološkome parseru HUMOR-u odnosno sintaktičkome parseru HumorESK-u, programi tvrtke MorphoLogic, DigiMorf, Helyësen i Elragoz. Iako su navedeni programi kreirani da odgovore posebnostima aglutinativnih jezika, u ovome će se odlomku prikazati neke od mogućnosti koje oni nude s ciljem poticanja razvijanja sličnih alata i softvera za hrvatski jezik[110].

[110] Opisi programa citirani su s internetske stranice http://www.morphologic.hu/index.php?option=com_virtuemart&page=shop.browse&category_id=59&Itemid=386, 22. srpnja 2008.

DigiMorf autori naime opisuju kao digitalni rječnik oblika riječi koji se temelji na sedam tomova *Rječnika mađarskoga jezika*. Program naime nudi dvije mogućnosti, pretraživanje rječnika i biranje riječi. Pri pretraživanju rječnika on daje informacije o složenosti neke riječi, vrsti riječi (za svaki pojedini oblik), tipu korijena riječi, paradigmi i posebnosti paradigme te valenciji glagola. Kod biranja riječi funkcija nudi dvanaest mogućnosti primjerice pretraživanje prema početku riječi, sredini riječi, kraju riječi, zatim prema sastavu riječi (broju korijena, afiksa), prema vrstama riječi (može se izabrati više vrsta riječi) te valenciji glagola.

Helyësen je program za uvježbavanje bilježenja mađarskoga zatvorenoga glasa *ë*, te nudi između ostaloga mogućnost prenošenja bilo kojega teksta u tekst s oznakama za navedeni glas. U slučaju višeznačnih riječi bira se najvjerojatnije rješenje, ali pritiskom na desnu tipku miša program nudi pregled konteksta te na taj način olakšava donošenje odluka.

Program Elragoz nudi pak pregled paradigmi svih mađarskih promjenjivih riječi, odnosno sve njihove oblike te mogućnost uvježbavanja njihovih oblika. Nudi isto tako osnovne morfološke informacije o riječima te mogućnost skupljanja riječi prema definiranim parametrima. Osim toga, program nudi i dinamičnu obradu rječničkih unosa, odnosno mogućnost pretvaranja nekoga oblika riječi u lemu ako taj oblik ima svoju paradigmu. Time se broj lema u rječničkoj bazi programa povećava za vrijeme njegova rada.
Zanimljiva je činjenica da se u slučaju navedenih programa za učenje mađarskoga jezika ne radi samo o softveru za učenje i poučavanje mađarskoga kao stranoga jezika. Iako se programi mogu rabiti i u te svrhe, količinom informacija koje nude, programima se mogu uspješno koristiti i izvorni govornici mađarskoga jezika.

9.2. Programi za strojno prevođenje

Morfološka se analiza, osim za stvaranje već spomenutih rječnika za učenje i poučavanje materinskoga i stranih jezika, može primijeniti i u različitim sustavima za prevođenje, odnosno u računalnim programima koji su u mogućnosti samostalno ponuditi prijevode kako pojedinih lema, tako i cijelih rečenica. U ovome će odlomku biti riječi o primjeni dostignuća na području morfološke analize jezika u različitim prevoditeljskim sustavima.

Cilj je ovoga prikaza osim davanja informacija o načinu rada različitih prevoditeljskih sustava dati i prikaz problema koji se pojavljuju pri izradi ovakvih i sličnih pomagala. Iz razloga što se prikazani problemi mogu djelomice usporediti s problemima pri izradi morfološkoga analizatora hrvatskoga standardnog jezika, smatram potrebnim objasniti problematiku analize jezika i sa stajališta izrade prevoditeljskih softvera. Razvijanje različitih pomagala za hrvatski jezik (među njima i prevoditeljskih softvera) jedno je od ključnih tema i hrvatske računalne lingvistike te se stoga odlučilo jedan odlomak posvetiti i informiranju o tome.

Istraživanja o izradi različitih sustava za prevođenje započela su već prije pedesetak godina, s razvojem informatičke tehnologije. Već prije nekoliko desetljeća stvoreni su sustavi za prevođenje koji su bili u mogućnosti ponuditi prihvatljivi rad, ali uz značajna ograničenja. Zanimljivo je međutim da, iako informatička tehnologija napreduje, današnji se prevoditeljsku sustavi grade na programima iz sedamdesetih godina prošloga stoljeća (Prószéky 2003. 9). Ipak, sustavi koji bi u potpunosti zamijenili prevoditelja do danas nisu razvijeni. Prema Prószékyjevim riječima:

> Istraživanja su dala mnoge rezultate, ali prevoditeljski sustav koji bi bez nadzora mogao u prihvatljivoj kvaliteti ponuditi prijevod teksta s nekoga drugog jezika do danas nije stvoren. Djelomični su nam rezultati, međutim, omogućili da napravimo takva pomagala koja će pokraj *ljudskoga prijevoda* – ako i neće moći obaviti posao umjesto prevoditelja – znatno ubrzati (i olakšati) prevođenje[111] (Prószéky 1999: 221).

Teorijski su moguća dva načina izrade prevoditeljskih sustava. U jednome se načinu prijevod rečenica nekoga teksta dobije korištenjem složenih programa za morfološki i sintaktički parsing te se uz pomoć dodatnih algoritama nudi odgovarajući prijevod, dok se u drugome slučaju, prema riječima stručnjaka, pribjeglo rješenju koje nudi „ekstenzivan razvoj informatičke tehnologije“. Kvalitetu, naime, zamjenjuje kvantiteta, odnosno na scenu stupaju prevoditeljske memorije koje iz velike baze podataka jednostavno pronalaze odgovarajući prijevod i nude ga korisniku (Prószéky 2003: 10). Ako se prevoditeljska memorija pak – slično kao baze podataka – podijeli među korisnicima, na taj način svaki

[111] Prijevod MA. Izvornik: „A kutatás számos részeredménnyel járt, de olyan számítógépes fordítórendszer, amely felügyelet nélkül elfogadható minőségben előállítja egy szöveg másik nyelvbeli fordítását, tulajdonképpen a mai napig nem született.1 A részeredmények azonban lehetővé teszik, hogy az emberi fordításhoz olyan számítógépes segédeszközöket készítsünk, amelyek – ha nem is végzik el a munkát a fordító helyett – jelentősen meggyorsítják (és megkönnyítik) a fordítást.“

prevoditelj na jednome poslužitelju raspolaže istom prevoditeljskom memorijom. Sustav tada iz iste baze podataka crpi različite prijevode, tako da se time poboljšava cjelokupna kvaliteta prijevoda (Prószéky 1999: 222). Osim toga, svaki prevoditelj ima mogućnost nadopunjavanja baze podataka odgovarajućim inputom – prijevodima, te na taj način korisnici intraneta uvelike poboljšavaju svoj rad.

Međutim, jedan od sustava koji se temelji na morfološkoj analizi i sintezi te koji omogućuje najbolje korištenje dosadašnjih dostignuća jest sustav MetaMorpho[112]. Prema riječima tvoraca ovoga prevoditeljskoga programa sustav se ne može nazvati ni sustavom koji radi prema načelu transfera, ni posrednika. Pri njegovu se radu koristi morfološka analiza i generiranje jer pokraj svakoga pravila kojim se koristi analizator nalazi se njegov par – pravilo kojim se koristi generator oblika. Tvorci su sustava odustali od zamisli metajezika te su za svaki jezik koji sudjeluje u prevođenju napravili poseban jezični modul. Time se po njihovim riječima poboljšava kvaliteta jer se rječnici mogu koristiti na prirodan način, ljudski se prijevodi mogu lako integrirati, lako se omogućuje proširivanje sustava uslijed djelovanja korisnika, a gramatika mora rješavati samo probleme aktualnoga jezičnog para[113].

Izradom morfološkoga analizatora hrvatskoga standardnog jezika uvelike se povećavaju mogućnosti za integriranje hrvatskoga jezika u sustav ovoga ili sličnih prevoditeljskih pomagala. Naime, sustav morfološke i sintaktičke je analize sadržan u formalizmu kojim se koristi MetaMorpho, a koji se naziva MMD format (MetaMorpho Dictionary). Svaki pojedini redak ili pravilo sadržava i dio koji se odnosi na analizu i dio koji se odnosi na generiranje. Za razliku od lema u rječniku, MMD sadrži jezične simbole te prikaz pravila na dva različita stupnja. Sintaktička se analiza događa uz pomoć sintaktičkoga parsera HumorESK-a (High-speed Unification based Morphology Enhanced with Syntactic Knowledge), koji je u mogućnosti ne samo bilježiti pojedine strukture i prikazati ih u grafičkome obliku, nego svaki pojedini tekst može transformirati bilo u koji formalizam – od kojih je jedna mogućnost njegov prijevod na drugi jezik (Prószéky-Kis 2002: 289).

[112] Opis sustava citiran je s internetske stranice http://www.morphologic.hu/index.php?option=com_content&task=view&id=433&Itemid=256), 22. sprnja 2008.

[113] http://www.morphologic.hu/index.php?option=com_content&task=view&id=433&Itemid=256), 22. sprnja 2008.

10. Završna riječ i perspektive

Nakon predočene problematike i rješenja u ovome će se poglavlju dati zaključci i nagovijestiti neki od budućih projekata na polju računalne obrade jezika. Iz razloga što se ovi projekti nalaze tek u fazi planiranja, smatram važnim predočiti i sažeti najvažnije ideje koje su se nametnule tijekom izrade ovoga rada kako s teorijskoga, tako i praktičnoga stajališta.

Praktična potreba koja se nametnula tijekom izrade morfološkoga analizatora za hrvatski standardni jezik, najviše se uočava na polju učenja i poučavanja hrvatskoga kao drugoga ili stranoga jezika. Naime, kao što je već prikazano, tijekom opisivanja hrvatskoga jezičnog sustava i njegova prilagođavanja formalizmu HUMOR-a pojavila se problematika opisa jezičnih zakonitosti u gramatikama i jezičnim priručnicima. Analogno tomu, ako se pretpostavi da učenik stranoga jezika, makar na određenom stupnju učenja jezika zaključuje slično računalu, odnosno kada se pomnije analiziraju problemi koji su se pojavili pri izradi morfološkoga parsera, dolazi se do zaključka da je potrebno razviti materijale koji će govornicima kojima hrvatski nije materinski jezik olakšavati ovladavanje jezičnim zakonitostima. Internom je anketom među profesorima hrvatskoga jezika koji su zaposleni u ustanovama koje polaze govornici kojima hrvatski nije materinski jezik zaključeno da opisana problematika nije samo teoretske prirode, nego da se pojavljuje i u praksi. Ako se tomu doda činjenica da je znanje hrvatskoga jezika sve potrebnije u pojedinim gospodarskim sferama te da se povećava broj govornika kojima hrvatski nije materinski jezik, a koji ga uče na različitim stupnjevima, opravdana je potreba za razvijanjem dodatnih materijala za njegovo što bolje svladavanje. Među te se materijale ubrajaju naime rječnici, jezični priručnici za govornike kojima hrvatski nije materinski jezik te prevoditeljski softver.

Prvim su se izborom kod razvijanja navedenih materijala postavili govornici njemačkoga jezika iz razloga što se najveći broj hrvatskih iseljenika nalazi na njemačkome govornom području. Druga bi ciljna skupina govornika bili govornici na mađarskome govornom području iz razloga što prema broju Hrvata u pojedinim državama Mađarska zauzima

sedmo mjesto[114]. Ako se naime uzme u obzir činjenica da je iseljavanje Hrvata na njemačko govorno područje započelo 1918. godine[115] te činjenica da se Hrvati u Mađarskoj smatraju autohtonom etničkom skupinom[116], da većina učenika pripadnika mađarske nacionalne manjine u Republici Hrvatskoj ima poteškoća sa služenjem hrvatskim jezikom[117], razvijanje projekta naznačenog u ovome radu u obliku njegove prilagodbe govornicima i mađarskoga jezika, može se smatrati potrebitim. Razvijanje softvera koji odgovara najzahtjevnijim korisnicima i suvremenim standardima koji su predstavljeni u poglavlju 9, smatra se prioritetom na polju rada hrvatske računalne lingvistike.

Razvijanje programa popraćeno je formalizmom opisa sustava hrvatskoga standardnog jezika te isto tako povlači za sobom problem izbora jezične varijante. U radu su se samo naznačili problemi, međutim, njihovo će se rješavanje morati odgoditi do jednoga od budućih projekata. Osim pitanja izbora jezične varijante, mnoga rješenja do kojih se došlo uslijed obrade različite problematike mogu pridonijeti točnijoj prezentaciji jezičnih zakonitosti te prilagodbi jezičnih opisa govornicima kojima hrvatski nije materinski jezik. To se prvenstveno odnosi na formalizaciju paradigmi promjenjivih vrsta riječi, promjenu strukture rječničkih unosa u rječnicima za učenike hrvatskoga kao drugoga ili stranoga jezika isticanjem uzorka paradigme i računalnoga roda. Osim toga, rječnički bi se oblici u učeničkim rječnicima trebali ustanoviti te bi se primjerice kod pridjeva koji imaju određeni i neodređeni oblik, radi lakšega korištenja uzorka paradigme, kao rječnički oblik mogao koristiti određeni, kao što je objašnjeno u poglavlju 7.3.

Iz razloga što je opseg ovoga rada ograničen, nije se moglo usredotočiti na sve probleme, nego su se istaknula samo ona područja koja su od važnosti za daljnja istraživanja i daljnje oblike primjene računalne morfološke analize hrvatskoga standardnog jezika. Točnije

[114] Države koje se nalaze prema broju iseljenika između Njemačke i Mađarske većinom su države s engleskoga govornog područja, za čije govornike postoje već objavljeni različiti nastavni materijali (Cvikić 2005b).
[115] http://www.mvpei.hr/hmiu/tekst.asp?q=osi001, 4. veljače 2008.
[116] http://www.mvpei.hr/hmiu/tekst.asp?q=hnm002, 4. veljače 2008.
[117] Ova se izjava temelji na anketi provedenoj među profesorima hrvatskoga jezika u Prosvjetno-kulturnom centru Mađara u Republici Hrvatskoj u Osijeku i sociolingvističkom istraživanju koje sam 2004. godine provela među učenicima navedene škole. Rezultati su istraživanja pokazali da su učenici, pripadnici mađarske nacionalne manjine, u omjeru 85% dvojezični govornici s izrazitom dominacijom mađarskoga jezika te da ih se svega 24% u svakodnevnoj interakciji koristi hrvatskim jezikom (Aleksa 2007). Prema subjektivnoj procjeni profesora hrvatskoga jezika u PKCM-u, hrvatska bi se jezična kompetencija većine mađarskih učenika mogla odrediti kao razina B1 i B2 Europskoga referentnog okvira. Za određivanje njihove stvarne jezične kompetencije i definiranje poteškoća i problema u svladavanju hrvatskoga jezika te postizanju viših razina znanja potrebna su dodatna istraživanja.

rečeno usredotočilo se samo na jezičnu sferu morfološkoga parsinga. Iz istoga se razloga predočio samo nacrt hrvatskoga računalnog morfološkog rječnika za govornike njemačkoga jezika. Postavljanjem temelja morfološkom rječniku u ovome radu ostvarile su se mogućnosti za njegovu kompilaciju koja bi se uskoro trebala i izraditi te time označiti nastavak ovoga i dosadašnjih sličnih projekata. Nastavak istraživanja isto tako podrazumijeva razvijanje dodatnih materijala ne samo na području translatologije i usvajanja jezika, nego i frazeologije, odnosno morfološke analize za izradu rječnika kolokacija hrvatskoga jezika, o čemu je bilo govora na različitim znanstvenim skupovima te o čemu će biti još riječi u nekim od sljedećih radova.

11. Összefoglaló

A horvát sztenderd nyelv számítógépes morfológiai elemzésének alkalmazási lehetőségei német nyelvterületen

1. Bevezetés

Jelen dolgozat a sztenderd horvát nyelvre összpontosít, ugyanakkor párhuzamosan tárgyal egy másik nyelvcsaládból származó nyelvet, a németet. Mindehhez kiindulópontként a nyelvek morfológiai elemzését veszi alapul. A téma megközelítése előtt fontos azonban bizonyos fogalmakat tisztázni, illetve néhány kérdésre választ adni. Az egyik a morfológiai elemzés választásának indokát taglalja, a második pedig a német nyelvre mint kitűzött célcsoportra vonatkozik.

1.1. Miért esett a döntés egy horvát morfológiai elemző kidolgozására?

Az Európai Tanács által létrehozott *Közös Európai Referenciakeret*[118] megjelenése után számottevő igény keletkezett a piacon olyan tananyagok kidolgozására, amelyek a horvát mint második és idegen nyelv tanítását segítik, mindamellett összhangban lesznek a Referenciakeretben leírt útmutatóval. A Referenciakeret horvát nyelvű kiadását részletesebben elemezve láthatjuk, hogy a nyelvtani kompetencia leírásánál a hangsúly többnyire a hagyományos nyelvtani szintekre kerül (Jelaska 2005a: 22). Ezt a tényt támasztja alá néhány Horvátországban elkezdett, illetve már befejezett projekt is, amelyek közül említést érdemel a *Hrvatski kao strani jezik: razvojna gramatika i rječnik* (MZOŠ 130738)[119] [A horvát mint idegen nyelv: nyelvtani fejlődés és szókincs] és a *Hrvatski kao drugi i strani jezik* (MZOŠ 0130438) [A horvát mint második és idegen nyelv] munkák. A

[118] Council of Europe (2001). *Common European Framework of Reference for Languages*: *learning, teaching, assessment*. Cambridge: Cambridge University Press, a továbbiakban „Referenciakeret"
[119] http://www.croatiana.org/croatiana--projekti-hrvatskikaostrani.htm, 2008 február 4.

projektek eredményei különböző tananyagok, egy morfológiai szótár és igeragozási útmutató, valamint a horvát alapszókincshez összegyűjtött, 700 szóból álló lista[120]. Mindezen eredmények már 2001 óta mutatják a projektek célját, azaz, hogy olyan tananyagok készüljenek, amelyek a horvát mint második és idegen nyelvet tanulók nyelvtani kompetenciájának fejlődését segítik elő. Megerősítésként a horvátot mint idegen nyelvet oktatók véleménye szolgál, akik többek között azt állítják, hogy a „meglévő nyelvtankönyvek nem felelnek meg sem az oktatók, sem a diákok igényeinek"[121](Cvikić 2005a: 320). Többször is hangsúlyozták, hogy szükség lenne nagyobb mennyiségű horvát mint idegen nyelv tanítását segítő taneszköz, nyelvtan és nyelvtani gyakorlókönyv, valamint az általános és szaknyelvet feldolgozó tankönyv, szótár, házi olvasmány kiadására (Cvikić 2005a: 320). Logikus folytatásként olyan, a horvátot mint második és idegen nyelvet segítő taneszközök megtervezését kezdeményeztük e dolgozat keretein belül, amelyek a horvát sztenderd nyelv automatikus morfológiai elemzésének felhasználásával valósíthatók meg.

A dolgozat a számítógépes morfológiai elemző két alkalmazási lehetőségét tárgyalja. Az első egy nyelvészeti jellegű alkalmazás, vagyis a nyelvtani paradigmák konkrét szemléltetése német anyanyelvűek számára. Ez az alkalmazás tulajdonképpen egy — a közeljövőben kidolgozandó — horvát számítógépes morfológiai szótár. Mindazonáltal az elemző révén létrehozott formalizmus a paradigmák újszerű megközelítését és elsajátítását mutatja be. A morfológiai szótár sajátossága, hogy a többi, már meglévő szótárhoz képest magában foglalja a Vladimir Anić *Veliki rječnik hrvatskoga jezika*[122] [Nagy horvát értelmező szótár] összes lemmáját, és nem csak nyelvtani leírásokon alapul, hanem korpuszelemzés, illetve a nyelvhasználat által pontosított leírásokat is tartalmaz.

A morfológiai elemzés második alkalmazási lehetősége a számítógépes fordítást illetve a különböző nyelvtanulást segítő eszközök létrehozása, amelyek horvát, német és magyar[123] viszonylatban is megvalósíthatók (vö. 9. fejezet). Mivel jelen dolgozat célja azon nyelvészeti problémák és dilemmák szemléltetése, amelyek a morfológiai elemző

[120] A pontos adatok és a kiadványok horvát megnevezése a dolgozatban olvasható.
[121] Eredeti szöveg: „Postojeće gramatike za izvorne govornike ne odgovaraju ni potrebama predavača ni potrebama učenika." (Fordítás MA).
[122] Anić, V. (2000). *Veliki rječnik hrvatskoga jezika*. Zagreb: Novi Liber, a továbbiakban RHJ.
[123] A magyar nyelvre való taneszközök kidolgozására Barić (2006) tanulmánya alapján nagy igény mutatkozik.

kidolgozása közben jelentkeztek, az elemző további alkalmazási lehetőségei csupán elméleti szinten kerülnek ábrázolásra.

1.2. Miért esett a választás a német anyanyelvűekre?

A fent említett morfológiai elemzésen alapuló eszközök elkészítése egy konkrét célcsoportot kívánt meg, ezért a német anyanyelvűekre a határon túli horvátok száma alapján esett a választás. A Horvát Külügyminisztérium adatai szerint (Prilog 1 — 1. függelék) a határon túli horvátok száma az angol és a német területen a legmagasabb. Figyelembe kell venni azt a tényt is, hogy a horvátok kivándorlása a német nyelvterületre már 1918-ban megkezdődött[124], illetve hogy már léteznek hasonló taneszközök angol anyanyelvűek számára, valamint azt a hipotézist, hogy a nem horvát anyanyelvűek számára[125] a legnagyobb gondot a horvát nyelvtan helyes elsajátítása jelenti. Többek között Németországban is számos helyen oktatják a horvátot második illetve idegen nyelvként (Jelaska 2005c). Mindezek alapján megalapozottnak látszik a fent említett eszközök kifejlesztésekor a német anyanyelvűekkel kezdeni. Mivel tudomásom szerint nem létezik az elvárásoknak megfelelő segédeszköz erre a célcsoportra, jelen dolgozat keretein belül született a már fent említett döntés.

1.3. A morfológiai elemzési eljárás kiválasztásának indokai

A morfológiai elemzés kiindulópontként való kiválasztásához két ok vezetett. Először egy olyan eljárást kellett kidolgozni, amely produktuma könnyen átalakítható és

[124] http://www.mvpei.hr/hmiu/tekst.asp?q=osi001, 2008 február 4.

[125] A hipotézis egy kérdőíves felmérésen alapul, amelyet a Horvátországi Oktatási és Művelődési Központ horvát szakos tanárai töltöttek ki, továbbá egy szociolingvisztikai vizsgálaton is alapul, amelyet 2004-ben végzett a szerző. A kutatás eredményei azt mutatják, hogy a tanulók 85%-a magyar anyanyelvű, és csak 24% használja a horvátot a mindennapi kommunikációban (Aleksa 2007a). A Horvátországi Oktatási és Művelődési Központ tanárai személyes véleménye szerint a tanulók többsége B1-es és B2-es nyelvi kompetencia szinttel rendelkezik. A kompetenciaszint pontos megállapításához azonban további kutatások szükségesek. E dolgozat keretein belül a kidolgozandó anyagok ezért elsősorban a B1 és B2, valamint a magasabb nyelvi szinteket célozzák meg.

implementálható bármilyen, nemcsak fordítást támogató rendszerbe, hanem más, a nyelvelsajátítást segítő eszközbe is. A választás az unifikációs alapú HUMOR[126] rendszerre esett. A második indok magában az eljárás előnyeiben rejlett, pontosabban abban a tényben, hogy az elemzésen alapuló eszközök olyan paradigmákat és nyelvtani információt tartalmaznak, amelyek korpuszvizsgálat és tényleges nyelvhasználat alapján lettek kidolgozva. Ez a szintézisen alapuló eszközökről nem minden esetben állítható (vö. 2. fejezet).

2. Horvát nyelvű tananyagok vizsgálata

Mivel e dolgozat mellékcélja a horvát morfológiai szótár kidolgozásához szükséges lépések leírása, ezért a horvát nyelvű tananyagok vizsgálata arra szolgál, hogy átvilágítsa a horvátot oktató tananyagokban szereplő szókincset. A vizsgálat betekintést nyújt a szókincs használatába, elemzi a nyelvtan és segédkönyvekben szereplő leírások pontosságát, egyértelműségét, és a horvátot mint második illetve idegen nyelvet tanulók[127] szempontjából való használhatóságát.

2.1. A tananyagok vizsgálatának elvei

A tananyagok vizsgálata két részletben történt meg. Az első részben — a dolgozat céljainak megfelelően — a német nyelvterületen használt német tankönyvek elemzése szerepel, a másodikban pedig a nyelvtan- és segédkönyvek vizsgálata történik. Mindkettő a *Kurikulum hrvatske nastave u inozemstvu*[128] [A horvát nyelv külföldön való oktatásának keretterve] elveit veszi figyelembe.

[126] High-speed Unification Based Morphology, a MorphoLogic által kidolgozott morfológiai elemzőrendszer.
[127] A fent említettekhez méltóan a tanulók alatt e dolgozatban a közép- illetve magasabb szintű nyelvi kompetenciával rendelkező diákok értendők.
[128] A *Kurikulum hrvatske nastave u inozemstvu* a horvátországi Oktatási Minisztérium honlapjáról származik: http://public.mzos.hr/Default.aspx?sec=2116, 2008. június 6., a továbbiakban: Kurikulum.

A dolgozat célkitűzéseinek megfelelően a tankönyvek vizsgálata a következő kérdésekre kísérel meg választ adni:

1. Mely szavak és szófajok szerepelnek leggyakrabban az elemzett tankönyvekben, valamint ezek melyik helyen találhatók a horvát nemzeti korpuszon alapuló gyakorisági szótárban?
2. Milyen nyelvtani és lexikai kollokációk (vö. 2.1.3. fejezet) fordulnak elő leggyakrabban az elemzett tankönyvekben, és ezek hogyan viszonyulnak gyakorisági szempontból a referens korpusz (vö. 6.3. fejezet) vizsgálatának eredményeihez?

A nyelvtan- és segédkönyvek vizsgálata a következő kérdésekre próbál válaszolni:

1. Milyen mértékben járul hozzá a paradigmák és a morfológiai leírások szemléltetése, továbbá a tankönyvekben leggyakrabban előforduló ragozható szavak és kollokációs egységek ragozási paradigmáinak leírása a horvát nyelv sikeres elsajátításához? A kérdést a következő szempontok alapján válaszoljuk meg:
    a) a leírások egyértelműsége;
    b) a leírások alkalmazhatósága.
2. Megfelelő információt nyújtanak-e a nyelvtankönyvek a tankönyvekben leggyakrabban előforduló szavak paradigmáit illetően?

A feltett kérdésekre korpuszvizsgálat segítségével adtuk meg a választ.

2.1.1. Korpuszvizsgálat

A szógyakoriság kiszámítására a *dtSearch*[129] programot használtuk, illetve a kollokációk vizsgálatát az *Ngram Statistics Package*[130] segítségével bonyolítottuk le. A továbbiakban a két program rövid ismertetése következik.

[129] http://www.dtsearch.com/customDevelop.html, 2008. április 29.

2.1.2. dtSearch

A szavak gyakorisági vizsgálatához a dtSearch program bizonyult a leghatékonyabbnak. A korpusz gyors indexelése révén statisztikai vizsgálatokra is megoldást kínál. Hiányosságai közé tartozik az, hogy a horvát szövegek lemmatizációja egyelőre nincs megoldva. Emiatt egy olyan horvát lemmatizálót kellett használni, amelyet a szerző a MorphoLogic eszközeinek segítségével dolgozott ki.

2.1.3. NSP

NSP — azaz az Ngram Statistics Package — egy, a Perl nyelven alapuló, Ted Pedersen kutatócsoportja által kialakított, ingyenesen hozzáférhető számítógépes programcsomag, amelynek segítségével nagy terjedelmű korpuszok vizsgálhatók. Összetett matematikai eljárások alapján számos hasznos információt nyújt a szókapcsolatokról, amely nagy segítséget nyújthat a nyelvészeti kutatómunka különböző területein (pl. nyelvtanítás, gyakorisági és frazeológiai szótárak szerkesztése).

A program elsősorban nem csak közvetlen szókapcsolatokat fedez fel egy elméletileg korlátlan mennyiségű szavakból álló szövegben, hanem olyan két vagy több szó közötti kapcsolatokat is vizsgál, amelyekben a szavakat esetlegesen néhány nyelvi egység (szó vagy mondattani jel) választja el egymástól. Az NSP működésének leírását összegezve azonban két dolog mindenképpen megemlítendő. Elsőként az eredmények relevánssága szempontjából a szókapcsolatok keresését lemmatizált szövegeken végezzük el, és csak ezután kezdjük el a kollokációkeresést[131]. Az ilyen módon kinyert eredmények így a

[130] http://www.d.umn.edu/~tpederse/nsp.html, 2008. április 29.

[131] A kollokáció mint fogalom meghatározása során több nyelvből származó különböző definícióval találkozhatunk. A Bakos Ferenc (2002: 338) *Idegen szavak és kifejezések szótár*ának definíciója szerint a kollokáció „gyakran együtt használt szavak szókapcsolata; állandósult szószerkezet / szókapcsolat". A német Duden szótár a kollokációt a nyelvi elemek tartalmi kombinálhatóságának, illetve különböző tartalmak egy logikai egységen belüli összeadásának tekinti. Ezzel szemben a horvát értelmező szótárból származó definíció alapján a kollokáció egy, a szavak közötti kötelező, nyelvtan által nem meghatározott szókapcsolat. Jelen munka a kollokációt egy, a nyelvtanilag nem meghatározott szókapcsolatnak tekinti, amely két vagy több szó között alakul ki, és ahol nagy valószínűséggel fennáll a lehetősége annak, hogy összetevői (a lexikai egységek) mindig egymás mellett vagy egymás közelében helyezkednek el.

tényleges nyelvhasználatot tükrözik. Ugyanis a szótári alakok kapcsolatát vizsgálva a kollokációk gyakoriságának és előfordulási valószínűségének kiszámítása relevánssá válik. Mivel szabadon elérhető horvát nyelvű lemmatizáló programok számomra nem ismertek, e dolgozat megírásánál a korábban említett horvát lemmatizálót használtam fel.

2.2. A német nyelvterületen fellelhető horvát tankönyvek vizsgálata

Jelen fejezet célja a 2.1. pontban megtárgyalt kérdések alapján, korpuszvizsgálat segítségével elemezni a német nyelvterületen található nyelvkönyvekben lévő szókincset, valamint az eredményeket összehasonlítani az alapszótár összeállítási elveivel. A vizsgálat a továbbiakban megkísérel kitekinteni azokra a nyelvtani információkra, amelyek szükségesek a német nyelvterületen fellelhető horvát tankönyvekből tanuló diákok számára, és amelyek nem találhatók meg a Kurikulum által definiált kiadványokban. A korpuszt két horvát nyelvű, német nyelvterületen használatos könyv alkotja, amelyeket eddig még nem vontak be hasonló kvantitatív és kvalitatív jellegű korpuszvizsgálatba (Blagus 2005a és 2005b, Cvikić 2005b: 221): a Drilo, S. *Kroatisch, TL 1, Lehrbuch für Anfänger* (2006)[132] c. tankönyve, valamint a *Kroatisch Lehrbuch für Fortgeschrittene* (1996)[133] c. kiadványa.

A vizsgálat eredményei a következők: a KA és KF összesen 7591 különböző szót tartalmaz, amelyek 3833 lemmára vezethetők vissza. A dolgozatban közölt 1. és 2. ábra a tankönyvekben fellelhető szófajok és szavak felosztását szemlélteti. A 3. ábra azokat a lemmákat tartalmazza, amelyek megjelennek a könyvekben, de ma már nem a horvát, hanem inkább vagy kizárólag a szerb szókészlet részei (erről bővebben a 6.2. fejezetben). A tankönyvek szókincse — néhány kivételtől eltekintve — összhangban van a horvát gyakorisági szótárban feltüntetett sorrenddel. Ha a morfológiai elemző, illetve szótár kidolgozásánál az elsőbbséget ezek a szavak kapják, a szöveg megértése és feldolgozása jelentősen javulhat. A *t-test* eredménye (Prilog 3. — 3. függelék) azt mutatja, hogy tokenekként a kollokációkban az igék szerepelnek a legnagyobb számban, utánuk a főnevek következnek. Mindazonáltal feltételezhetjük, hogy nagyobb igény lenne a nyelvtankönyvekben szereplő igei és főnévi paradigmára vonatkozó információk

[132] A továbbiakban: KA.
[133] A továbbiakban: KF.

kikeresésére. A *Pointwise Mutual Information* elemzés (Prilog 4 — 4. függelék) pedig a melléknévi paradigma fontosságára utal. Többek között azt bizonyítja, hogy a tényleges nyelvhasználatban gyakran fordulnak elő melléknévi kollokációk is. A szavak szintaktikai viszonyáról további magyarázattal az NSP által lebonyolított tri- és 4-gram elemzés szolgál, amelyet az 5. és 6. függelék tartalmaz. Ezen túlmenően a szavak szintaktikai tulajdonságainak leírása és elemzése jelen dolgozat keretein kívül esik.

2.3. A horvát nyelv oktatásában használt nyelvtankönyvek elemzése

Az elemzendő nyelvtankönyvek kiválasztása a horvát Oktatási Minisztérium által engedélyezett kiadványokon[134], valamint a horvát szakos tanárok (akik a horvátot nem horvát anyanyelvűeknek tanítják[135]) által kitöltött kérdőív válaszain alapul. A végső választás a reprezentativitás elv felhasználásával a következő kiadványokra esett:

- J. Silić–I. Pranjković (2005) *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*[136],
- E. Barić et al. (1995) *Hrvatska gramatika*[137],
- D. Raguž (1997) *Praktična hrvatska gramatika*[138],
- a *Gramatički tezaurus hrvatskoga jezika v 1.2.* c. elektronikus nyelvtani szótár[139],
- és a *Hrvatski morfološki leksikon* horvát morfológiai lexikon internetes verziója[140].

[134] Az elemzésbe bevett kiadványok a *Kurikulum hrvatske nastave u inozemstvu* c. dokumentum 8.1.3 pontja alatt vannak felsorolva (http://public.mzos.hr/Default.aspx?sec=2116-a, 2008. június 8.). Mivel a többi feltüntetett kiadványt nem használják a horvát nyelv tanítására a B1-es és B2-es szinteken, ezért az elemzésben nem szerepelnek.
[135] A dolgozat jellegzetességeit és a német nyelvterületen dolgozó horvát tanárok a szerző számára való elérhetetlenségét figyelembe véve, a kérdőívet csak olyan horvát szakos tanárok töltötték ki, akik magyar anyanyelvű diákoknak horvátot mint idegen nyelvet tanítanak.
[136] A továbbiakban: GHJ.
[137] A továbbiakban: HG.
[138] A továbbiakban: PHG.
[139] A továbbiakban: GT.
[140] A továbbiakban: HML.

2.3.1. Igék

Az igei paradigma vizsgálata után az a következtetés szűrhető le, hogy a horvátot mint második vagy idegen nyelvet tanuló diákok számára egy újszerű megközelítés szükséges, amely elsősorban a korpuszelemzésen alapul. Mindezt bővebben a 7.4.2. fejezet tárgyalja.

2.3.2. Melléknevek

A mellékneveket érintő leírások elemzésével a fentihez igen hasonló következtetésre juthattunk. Felmerül azonban a kérdés, hogy vajon mekkora a tényleges melléknévi lemmák száma a RHJ-ben, ugyanis a szótárban felsorolt melléknevek esetén következetlenségek vehetők észre (Aleksa 2007b). Az újszerű melléknévi paradigma szemléltetése a dolgozat 7.3. fejezetében tekinthető meg.

2.3.3. Főnevek

Ha az elemzett segéd- és nyelvtankönyvekben, valamint Anić (2000) szótárában tüzetesebben megvizsgáljuk a főnévi paradigmát, először a szükséges információk hiánya illetve nem egyértelműsége tűnik fel, utána pedig a grammatikai nem problémája, amelyről a 8.4.2. fejezetben lehet bővebben olvasni.

2.3.4. Névmások és számok

A nyelvtan- és segédkönyvekben szereplő névmások és számnevek paradigmájáról szóló információk vizsgálata után arra a következtetésre jutottunk, hogy a nem horvát anyanyelvűek számára egy formalizmus általi megközelítés konkrét információkat nyújthat, és minden morfológiával kapcsolatos kérdésre konkrét választ ad. A nyelvi

formalizmus a HUMOR leírásán alapul, amely tulajdonképpen a számítógépes morfológiai szótár alapja is.

2.3.5. A Vladimir Anić *Veliki rječnik hrvatskoga jezika* c. horvát értelmező szótár vizsgálata

A vizsgálat alapján az a következtetés vonható le, hogy a nem anyanyelvűek számára nem elegendők a RHJ-ban lévő információk, valamint az, hogy egy más jellegű tanulói szótár megszerkesztése kívánatos, amely — ahogyan az Prószéky (1997) tanulmányában olvasható — dinamikus keresési lehetőségeket nyújt. Egy ilyen jellegű szótár létrehozásához mindenképpen a jelen dolgozat által felvázolt morfológiai szótár tekinthető alapnak.

3. Morfológiai elemzés a morfológiai szótár kidolgozásának szolgálatában

3.1. Alapelvek a horvát sztenderd nyelv morfológiai szótárának kidolgozásához

A horvát morfológiai szótár kidolgozása előtt felmerült néhány kétely és kérdés, amelyeket szükséges volt megválaszolni. Először is, mivel a horvát számítógépes morfológiai szótár a horvát sztenderd nyelv morfológiai elemzésének egyik alkalmazási lehetősége, fontosnak bizonyult bővebben megindokolni a szótár kidolgozásához kiválasztott eljárást, tulajdonképpen az elemzés és nem a szintézis használatát.

A tankönyvek vizsgálata másfajta megközelítést tett szükségessé, ezért fontos volt másmilyen jellegű, a tényleges nyelvhasználathoz illeszkedő taneszközök kidolgozásáról elgondolkodni, amelyek eltérnek a hagyományos, szintézisen alapulóktól. Tadić (1994: 31) dolgozatában megemlíti azt a tényt, hogy szintézissel mint eljárással többnyire a kutatói hipotézisek tesztelhetők, és célja alapvetően nem az alakok felismerése. A horvát morfológiai elemző kidolgozása során, a szövegek közel 100%-os felismerésére törekedve, a lexikai adatbázis nemcsak a RHJ-ból származó szavakat tartalmazza, hanem a tesztkorpusz vizsgálatának eredményeit is figyelembe veszi. Ily módon egy kibővített,

nyelvhasználat által pontosított alakban alapot nyújt a morfológiai szótár kidolgozásához (vö. 4.2. fejezet).
A morfológiai szótárral kapcsolatban a második felmerülő kérdés a nyelvi változat kiválasztásának problémáját érinti.

3.2. A morfológiai szótár számára kiválasztott nyelvi változat

A horvát nyelvi változat kérdése elsősorban a sztenderd nyelvi változat korlátozására, illetve a szótárba az egyéb változatokból származó lemmák felvételére vonatkozik. Tulajdonképpen két lehetőség kínálkozik: az első a morfológiai elemző sztenderd nyelvi alap közvetlen felhasználása. A másik pedig, a német nyelvterületen található horvát tankönyvek elemzését kiindulópontként felhasználva (vö. 2.2. fejezet) a horvát nyelv egyéb változataiból származó lemmák feldolgozása.
A döntés végső soron csak a horvát sztenderd nyelvi változat felhasználására esett, mivel a Kurikulum a többi nyelvváltozatot csupán receptív készségek szintjén kezeli.

4. HUMOR mint egyfajta morfológiai elemző

A HUMOR, azaz a High-Speed Unification based Morphology olyan unifikációs alapú, a MorphoLogic által kidolgozott morfológiai elemző program, amelyet több nyelv morfológiai elemzésére fejlesztettek ki (Prószéky–Kis 1999a). A program univerzalitását sokoldalú használata is bizonyítja: a HUMOR az alapja például a MobiDic és MobiMouse szótárkezelő alkalmazásnak, valamint a MetaMorpho programnak, beépíthető különböző helyesírás ellenőrzőkbe és a nyelvek morfológiai elemzésére is kiválóan használható (vö. 9. fejezet). Demóverziója 1992-ben készült, és ma már a működő lengyel, német, angol, magyar, valamint számos, kidolgozás alatt álló változaton túl, a jelen dolgozat keretében a horvát változat is elkészült.

4.1. A HUMOR működése

Röviden összefoglalva a morfológiai elemzés a szavak stem-re és term-re való felbontásán alapul. Stem-nek tekintjük a szó azon részét, amelynek írása a ragozás során nem változik meg. Például a *noga* 'láb' szó esetében stem-nek a *no-* részt tekintjük, mivel a ragozás során megjelenik a *nozi* változat is, szótőnek ugyanakkor a *nog* tekintendő. Ugyanebből adódik az, hogy a horvát term, amely alatt egyszerűsítve a szó végződését értjük, hosszabb és változékonyabb, mint a hagyományosan értelmezett horvát toldalék. Mivel a HUMOR leírása nem a hagyományos grammatikai kategóriákkal operál, fontos volt a program kidolgozásánál új fogalmakat bevezetni a szótő, illetve a rag[141] helyett (Prószéky–Kis 1999a). A HUMOR részletes leírásáról, elemzési eljárásának működéséről, illetve a magyar és horvát verzió működésének jellegzetességeiről a dolgozat e fejezetében olvasható bővebben.

4.2. A morfológiai elemző struktúrája

A morfológiai elemző struktúrája több részre osztható. A technikai, ún. *engine* egységen kívül, még egy adatbázis is tartozik hozzá. Nyelvészeti szempontból a lexikai alap (a HUMOR-féle szótár) a legfontosabb. A horvát nyelv esetében az Anić (2000) szótárának teljes készletét magában foglalja. A lexikai alap azonban nem egy meghatározott, véges számú elemből áll össze, hanem szükség esetén bármikor új lemmákkal bővíthető.

4.3. A fogalmak definiálása

A morfológiai elemzés működésének részletesebb leírása előtt fontos néhány fogalmat meghatározni.

[141] Jelen dolgozatban a *rag* kifejezés alatt, ha másképpen nincs kiemelve, a horvát *nastavak*, illetve *sufiks* megfelelőjét értjük, ami megfelel a magyar *toldalék* fogalmának és magában foglalja a *képző*t, a *jel*et és a *rag*ot.

1. Mivel maga a rendszer a szóközt elemzési határnak tekinti, és a szót grafikus szinten értelmezi (pl. az összetett igeidők felismerése ez által problémát okozhat), a szócsoport elemei az első fázisban csak akkor ismerhetők fel, ha a lexikai adatbázisban lemmák részeként szerepelnek. Pl. a visszaható névmás *sebe/se* együtt szerepelhet a *kupati* ʻfürdetʼ igével a *kupati se* ʻfürdikʼ jelentésben is. Mivel a *kupati* ige nem vonz kötelezően visszaható névmást, az ige a lexikai adatbázisban így kétszer szerepel: *kupati*-ként, illetve *kupati se*-ként. Ezzel a szóköz problémája részben megoldható.

   Mivel a jelen tanulmány csak a morfológiai analízis problémáival foglalkozik, eltekintünk a szintaktikai, fonetikai és pragmatikai aspektustól. Mindazonáltal amennyiben a HumorESK-et, vagyis a HUMOR szintaktikai tudással gazdagított változatát használjuk fel különböző programokban, ezzel növelhetjük a szöveg számítógép általi felismerését.
2. A hagyományos nyelvtantól eltérően, a HUMOR rendszer keretein belül a szavak szófajhoz való tartozását a paradigmája határozza meg (vö. 7. fejezet). Pl. a melléknév kategóriája minden melléknévi paradigmával rendelkező szót tartalmaz.
3. A HUMOR rendszer keretein belül kidolgozott megoldások hasznosságát egy pilotkutatással tervezzük a jövőben felülvizsgálni. A morfológiai szótár kidolgozása jelenleg az utolsó fázisnál tart (vö. 8. fejezet).
4. Az elemzésnél a horvát összetett betűk (*dž, lj* és *nj*) egy egységnek tekintendők.

5. Az agglutináló, flektáló és magasan flektáló nyelvekre kidolgozott egyéb morfológiai elemzők áttekintése

A HUMOR szerzői (Prószéky–Kis 1999a: 261) többször is hangsúlyozták a rendszer kétszintű morfológiai unifikációs alapját (Koskenniemi 1983), amely a német GERTWOL leírásában is szerepel (Haapalainen–Majorin 1994). Ez a fejezet részletesen tárgyalja a német nyelv specifikumait és a GERTWOL által nyújtott megoldásokat, valamint összehasonlítja ezeket a horvát, magyar és lengyel nyelvleírási problémák jellegzetességeivel. Mivel a morfológiai elemző egyik jövőbeli alkalmazása a horvát–német, illetve horvát–magyar fordítóeszközök kidolgozása, jelen fejezet bővebb

információt nyújt a két rendszerről, és a fentiekben említett nyelvcsoportba tartozó nyelvek feldolgozásáról.

6. A horvát morfológiai elemző kidolgozása

6.1. A nyelvtörténet és nyelvpolitika kérdése

A modern horvát nyelv csak a XX. század 90-es éveiben alakult át egy erős nyelvtisztító folyamat révén, az 1990 előtti művek szerbhorvát nyelven íródtak. A sztenderd horvát nyelv kodifikálása 1835-ben kezdődött el, majd a folyamat megszakadása után a szerbhorvát nyelv kialakulása következett. Emiatt a horvát nyelvnek három fontosabb változata létezik: az 1835 előtti, a szerbhorvát nyelvi változat 1990-ig, illetve a modern horvát nyelv. Mivel a lengyel (a HUMOR egyetlen kidolgozott szláv nyelvi változata, amely a latin írást használja) lehetővé teszi a XVIII. századi szövegek elemzését is (Wołosz 2005), felmerül a kérdés, hogy a horvát változatnak is kell-e ilyen lehetőséget biztosítania. Ha az esetlegesen felmerülő igénynek eleget kívánunk tenni, a meglévő szókészlet jelentős bővítése válna szükségessé, és figyelembe kellene venni a különböző nyelvtani és helyesírási szabályokat, valamint a szerb nyelvből származó lexémákat, ami az adott esetben a program elnevezésében a horvát jelző használatát megkérdőjelezi. A végső döntés a horvát sztenderd nyelvi változatra esett a szókészlet bővítésének lehetőségével, ha a korpuszvizsgálat egy szónak illetve a szó változatának gyakori használatát mutatja, és ezáltal jelentősen javulhat a szövegfelismerés.

6.2. A nyelvváltozat kérdése

A 3.2. fejezetben tárgyaltakhoz hasonlóan, a morfológiai elemző kidolgozásánál is felmerült a nyelvváltozat kérdése. A dilemma az elemzőnek az írott változatra való korlátozása, és az esetleg beszélt változat felvétele között állt, valamint a helyi és egyéb nyelvi változatok tipikus lexémáinak hozzáadására vonatkozott. A végső döntés az elemző írott változatára esett, a többi változatból származó lexémák az elemző lexikai adatbázisába

való bevételi lehetőségének fenntartásával, amennyiben ennek szükségességét a korpuszvizsgálat alátámasztja.

6.3. A tesztkorpusz

Mivel számunkra a teljes Horvát Nemzeti Korpusz (HNK) nem volt elérhető, a rendszert egy, a HNK mintáján készült tesztkorpuszon tettük próbára. A korpusz öt alkorpuszból áll össze, amelyek a következő részekre oszthatók:

1. a legfontosabb horvát irodalmi művek (dráma, próza és líra);
2. efemer anyagok és nyomtatott írások;
3. elektronikus médiából származó szövegek (internetes oldalakról, blogokból);
4. szakirodalomból származó szövegek;
5. különböző tankönyvekből származó szövegek.

A műfajok kiválasztásának indokairól, illetve a korpusz összeállításának kérdéseiről bővebb információ a dolgozat e fejezetében található.

7. A morfológiai elemző használata a horvát sztenderd nyelv oktatására és tanulására

A 2.3.5. fejezetben leírtak szerint igény támadt a horvát nyelv oktatását és tanulását segítő különböző eszközök kidolgozására. A CALL (Computer-Aided Language Learning) (Prószéky 1997b) keretén belül, a Prószéky által összefoglalt igényeknek megfelelően, olyan eszközök kidolgozása vált szükségessé, amelyek a tanulóknak dinamikus módon, megfelelő információval szolgálnak szavak paradigmáiról, valamint összekötik a horvát lemmákat a tanulók anyanyelvéből származó lemmákkal is. Ennek következtében ebben a fejezetben, a jelen dolgozat célkitűzéseit figyelembe véve, a horvát nyelv számítógépes, német anyanyelvűek részére kidogozott morfológiai szótárának felvázolására kerül sor. A továbbiakban a felmerült kérdésekről és problémákról esik szó.

7.1. A horvát ragozható szófajok feldolgozásának problémái és ezek megoldási lehetőségei

A morfológiai elemző kidolgozása során több, a ragozható szófajt (horvátul *promjenjiva vrsta riječi*) érintő probléma keletkezett. A jelen dolgozat csak a legfontosabbakat szemlélteti, valamint azokat a megoldásokat mutatja be, amelyek jelentősen eltérnek a hagyományos nyelvtani reprezentálástól, és segítséget nyújtanak a horvát mint második és idegen nyelv oktatásában.

7.2. Főnevek

7.2.1. Főnévi paradigma

Ha a nem horvát anyanyelvű tanuló a nyelvtan- és segédkönyvekben szereplő információkra támaszkodik, az alábbi adatok szükségesek számára, hogy sikeresen generáljon egy helyes horvát főnévi paradigmát:

- az adott hímnemű főnév élő vagy élettelen fogalmat ábrázol-e;
    - melyik ragozásról van szó (*-a, -e* vagy *-i deklinacija*);[142]
    - a főnév alapja

A HUMOR rendszer által kialakított ragozási mátrixhoz (a dolgozat 24. ábrája) nem szükséges a fent említett információ, illetve eljárás. Csupán az alábbi, minden egyes főnévnél feltüntetett információ segítségével létrehozható a helyes főnévi paradigma:

- a szó stem-je
- az adott hímnemű főnév élő vagy élettelen kategóriához tartozik-e;

[142] A főnevek egyes ragozási csoportokba való besorolása a főnév genitív alakja szerint történik. Más szóval, a HG szerint a végződéseket ahhoz az alaphoz illesztjük, amelyet általában úgy állapítunk meg, hogy a genitív egyes számból elhagyjuk a végződést (Barić et al. 1995: 104). (A horvát szöveg: „nastavci se dodaju na osnovu, koja se, u pravilu, dobije ako se u gen. jedn. izostavi nastavak”.)

- a főnév 11-es (nominatív egyes szám), 12-es (genitív egyes szám), 21-es (nominatív többes szám) és 22-es (genitív többes szám) term-je

Így pl. a *more* 'tenger' vagy a *dječak* 'kisfiú' lemma paradigmájához szükséges információk a következők lennének. A függőleges vonal a stem-nek a term-től való elválasztását jelenti.

mor|e sr.r (nž., -e, -a, -a, -a)
dječa|k m.r (ž., -k, -ka, -ci, -ka)

Ha a pilotkutatás a 24. ábrában feltüntetett főnévi mátrix hasznosságát bizonyítja, szükségessé válik egy ilyen információval gazdagított tanulói szótárt is összeállítani.

7.2.2. Nyelvészeti problémák

A főnévi paradigmát illetően a legfontosabb nyelvészeti problémakörök közé az esetek száma és a vokatívusz eset meghatározása tartozik. Mivel a főnévi, továbbá a melléknévi paradigmánál a datívusz és lokatívusz esetek kivétel nélkül egyformának bizonyultak morfológiai szinten, a morfológiai szótárban hét eset helyett elegendő hatot ábrázolni. A vokatívusz problémáját, valamint az esetek további kérdéseit bővebben a 8.4.1-es fejezet tárgyalja.

7.3. Melléknevek

A hagyományos nyelvtani leírás szerint a horvát nyelvtan a mellékneveket először szemantikai tulajdonságuk, majd határozottsági tulajdonságuk alapján csoportosítja, majd három nemben ragozza (egyes és többes számban), élő vagy élettelenként, a melléknevet követő főnévtől függően. Ebből következik, hogy pl. a *crven* (piros) leíró melléknévnek 93 különböző, az azonosakkal együtt összesen 123 ún. ragozási cellája létezik. Fontos azonban hangsúlyozni, hogy az említett ragozási alakok csak az írott szöveg elemzésére

vonatkoznak. A beszélt nyelvben a hangsúlyt is figyelembe véve a különböző ragozási cellák száma tovább bővül. A HUMOR rendszer számára kidolgozott melléknévi mátrix (a dolgozat 27. ábrája) és minden melléknév tizenhat különböző alakja segítségével (a dolgozat 26. ábrája) létrehozható a teljes melléknévi paradigma. A melléknevek stem-jeik illetve term-jeik alapján 21 ragozási csoportba sorolhatók be (a dolgozat 30. ábrája). Döntőnek minősülnek az 1-es és az 5-ös (szótári) alakok, amelyek párban való előfordulása kizárólag a leíró melléknevek tulajdonsága. Kivételt képeznek a főnévből képzett melléknevek, amelyeknek a ragozási alakjai azonosak a leíró melléknevekéivel, viszont a paradigmájuk kizárja a határozott ragozást.

7.3.1. Melléknevek fokozása

Mivel a HUMOR leírásában a melléknevek közép- illetve felsőfokot képező ragjai különböző term-eket és stem-eket alkothatnának (ami jelentős mértékben bonyolítaná a melléknévi ragozás rendszerezését), így ezek a fokozott melléknevek külön lemmaként szerepelnek a szótárban, és természetesen utalnak az adott melléknév szótári alakjára.

7.3.2. Nyelvészeti problémák

A bemutatott melléknévi rendszer leírása nemcsak számítógépes kezelés szempontjából hasznos, hanem a nem anyanyelvűek számára is nagy segítséget nyújt a horvát nyelv elsajátításában, mert elsősorban nem a szavak szemantikai tulajdonságát veszi alapul, hanem azok morfológiai felépítését. Továbbá nincs szükség mind a 93 ragozási cella megtanulására, hanem csak a melléknév tizenhat különböző alakját kell elsajátítani.
Az első előbukkanó probléma a hím- és semlegesnemű alakkal nem rendelkező melléknevek ragozását (pl. *trudna* ʻterhes’), majd a második a nem fokozható melléknevek fokozását illeti. Ezeket a kérdéseket a dolgozatban alaposabban górcső alá vesszük. Gondot okoznak a helyesírási eltérések is. Például a *mađarski* ʻmagyar’ melléknévnek a Babić – Finka – Moguš (1995) *Hrvatski pravopis* c. horvát helyesírási szótár szerint két írásmódja létezik: a *mađarski* és a *madžarski* (1995: 277). Ezeket azonban Anić értelmező

szótára nem tünteti fel, és csupán az első említett alakot közli. Megoldásként mindkét alak külön lemmaként szerepel a HUMOR szótárában.

7.3.3. A horvát melléknévi paradigma számítógépes feldolgozásának összehasonlítása más nyelvek melléknévi paradigmájának feldolgozásával

A kidolgozott nyelvi formalizmus hasznosításának hasonló példái más szláv nyelvekben is megtalálhatók. A dolgozat 31. ábrája az orosz példát ábrázolja, ám hasonló megoldás a lengyel nyelvből is idézhető. Az említett nyelvekből származó szótári megoldások a horvát morfológiai szótár elkészítésénél is alkalmazhatók.

7.4. Igék

7.4.1. Igei paradigma

7.4.2. Igei paradigma a morfológiai elemzőn belül

A 2.3.1-es fejezetben lévő igei paradigma szemléltetése alapján arra a következtetésre juthatunk, hogy az igei paradigma a horvát nyelvtan egyik legtöbbet tárgyalt tétele. A szemléltetések és magyarázatok változatossága és ellentétes volta (vö. 2.3.1-es fejezet) arra utal, hogy szükség lenne egy új paradigmaleírásra, amely konkrét válaszokkal szolgál majd a tanulóknak. Anić (2000) szótárában 12 413 ige szerepel, amelyeket a hagyományos nyelvtan hét igeragozási csoportba, ezek alatt további osztályokra bontja. Az igeragozást illető nyelvészeti problémák elsősorban a múlt igeidők alakjainál (*aorist, imperfekt*), az igei főnév (*glagolska imenica*) és igei melléknév (*glagolski pridjev*) kategorizálásánál, valamint ezek főnévi, melléknévi illetve igei paradigmához való besorolásánál jelentkeznek. A HUMOR számára javasolt megoldás a korpuszvizsgálat alapján megpróbál ezekre a kérdésekre válaszolni. Az igéket stem-jeik és term-jeik alapján 96 ragozási csoportba sorolja. A csoportokat képviselő paradigmák az Anić (2000) szótárában meglévő összes ige helyes alakjának generálását biztosítják, miként az a 35. és a 36. ábráról leolvasható.

7.4.3. További problémák

Külön nehézséget okoz, hogy a HUMOR rendszer számára a szóköz vagy a szóköz értékű karakter mindig az elemzési határt jelenti. Az ebből adódó problémák a HumorESK esetleges horvát változatának használatával valamennyire orvosolhatók. A horvát morfológiai rendszer leírásánál figyelembe kellett venni azt is, hogy sok olyan szó létezik, amelynek ragozása egyedi, és nem követ egy-egy típust. Ezeket a szavakat mind külön lehetett csak a HUMOR szótárába felvenni.

8. A horvát sztenderd nyelv számítógépes morfológiai szótára

8.1. A szótár struktúrája

8.2. Technikai megoldások

8.3. A morfológiai szótár további alkalmazási lehetőségei

A felhasználók igényeit és a már érintett problémákat figyelembe véve, a tervezett morfológiai szótár több táblából álló összetett adatbázis szerkezetének köszönhetően a következő lehetőségeket nyújtaná:

1. A felhasználó kiválasztaná a keresett szót a horvát szólistából, vagy az idegen nyelvű listából, amely a szó horvát megfelelőjére automatikusan hivatkozna.
2. Az idegen nyelvű szólistán kívül a felhasználó választhatna több nézet közül. A lemmákat ábécé sorrendben soroltathatja fel, vagy *a tergo* sorrendben.
3. A lemma kiválasztása után a felhasználónak az egyes lemma, illetve paradigmára vonatkozó információk rendelkezésére állnának:
    a) „úgy ragozódik mint az x”
    b) az y-dik kategóriához tartozik
    c) szükséges nyelvtani információk (nem, szám, stb.)
    d) a szó változatlan részét (stem-et) úgy alkotjuk, hogy a szó végéből z-számú karaktert vágunk le, és ehhez adjuk utána a term-eket
    e) a teljes paradigma
    f) a lemma illetve a szóalak kontextuson belüli használata
    g) a szavak elválasztásának módja

A szótár kibővíthető különböző egyéb információkkal, mint pl. kollokációkkal. Ez azonban egy jövőbeli projekt célja.

8.4. Nyelvi kérdések és megoldások

8.4.1. Esetek

A szótár kidolgozása során több olyan kérdés merült fel, amelyre megfelelő megoldást kellett találni. Ezek a megoldások azonban megkérdőjelezik az eddigi hagyományt és egyes tételek újszerű megközelítését kívánják meg. Az elemző által használt formalizmus minden bizonnyal meg tudná könnyíteni a kérdéses tételek bemutatását és tanítását nem horvát anyanyelvű diákoknak. Az első kérdés, amely részletesebb kifejtést igényel, az esetek kérdése. A második pedig a grammatikai nemre vonatkozik.

A datívusz és a lokatívusz eseteken kívül az egyik legproblémásabb eset a vokatívusz (a probléma részletesebben a dolgozat e fejezetében olvasható). A nehézségek voltaképpen a nyelvtankönyvekben lévő információk mennyiségében és minőségében tükröződnek. Mivel az egyes szavak helyes vokatívusz alakját illető kérdésre még a korpuszvizsgálat sem tud elfogadható megoldást adni (pl. a vokatívusz eset alakja mindig egybeesik a paradigma valamelyik szóalakjával és ez által csak (félig) manuálisan emelhető ki a korpuszból), így a kérdés további kutatást igényel egy jövőbeli projekt keretein belül.

8.4.2. A nemek problémája

A szavak neme szintaktikai kategória, amit a Barić et al. (1995: 101) *Hrvatska gramatika* c. nyelvtankönyv is állít. Mivel a lengyel nyelvészek több nyelvtani nem létezését bizonyították a lengyelben (Saloni et al. 2007b: 9), felmerült a kérdés, hogy vajon a hagyományos horvát grammatikai nemek felhasználásával a diákok képesek lennének-e grammatikailag helyes mondatokat generálni. Mivel a válasz egyértelműen negatívnak bizonyult, több újszerű nyelvtani nem kategória létrehozása vált szükségessé, és egyben elkerülhetetlenné. Ezek a nemek a hagyományos nyelvtani nemek alkategóriáinak

tekinthetők, és segítségükkel a számítógép, illetve a nem horvát anyanyelvű tanulók képesek lennének helyes mondatokat generálni.
Ezen felfogás szerint kilenc nyelvtani nem létezik a horvát nyelvben, amelyet az alábbi mondatba a főnevek behelyezése bizonyít:

Iako [*biti*] ovdje [imenica] [odnosna zamjenica] volim, ne vidim [pokazna zamjenica]
Habár [*van*] itt [főnév] [vonatkozó névmás] szeretek, nem látom [mutató névmás]

[*dva*] [imenica].
[*kettő*] [főnév].

A mondatban használt számnév, illetve a vonatkozó névmás alakja bizonyítja az egyes nemek létezését, ahogyan az a dolgozat 44. ábrájáról is leolvasható.

9. A morfológiai elemzés további alkalmazása

9.1. Nyelvtanulásra és -oktatásra kidolgozott eszközök

9.1.1. Egynyelvű és többnyelvű szótárak

9.1.2. Magyar nyelv tanulására kidolgozott eszközök

9.2. Fordítástámogató rendszerek

A HUMOR rendszer unifikációs alapja lehetőséget kínál többféle rendszerbe való implementálásra. Jelen fejezetben a MorphoLogic által kidolgozott, a HUMOR illetve HumorESK rendszert használó eszközök bemutatására is sor került. Fontos azonban megemlíteni, hogy a bemutatott eszközök célja nem reklám, csupán szemléltető, informatív jellegű. A bemutatandó eszközök kiválasztásánál a szakmai elismerés döntött. Ebben a fejezetben a következő eszközök lettek bemutatva: MobiDic (*multi-dictionary environment)*, MobiMouse (*context-sensitive instant comprehension tool*), Słownik gramatyczny języka polskiego, Rječnik hrvatskoga jezika na CD-romu, DUDEN Großwörterbuch der deutschen Sprache, DigiMorf, Helyësen, Elragoz és a Metamorpho (*MetaMorpho Dictionary*). A fejezet részletesen kitér az eszközök előnyeire és hátrányaira is.

10. Kitekintés

Jelen dolgozat két szükségletre világított rá. Az egyik praktikus jellegű — a horvát mint második illetve idegen nyelv tanításához szükséges segédeszközök kidolgozását jelentette —, a másik pedig elméleti, és a horvát nyelvtanításban a nem horvát anyanyelvűeknek szóló tételek újszerű feldolgozását és ábrázolását foglalja magában. A dolgozat korlátait figyelembe véve nem minden probléma és megoldás került bemutatásra, a hangsúly csak a legfontosabb, újszerű nyelvészeti feldolgozás témakörébe tartozó tételekre helyeződött. A jövőbeli tervek közé tartozik a morfológiai elemzés egyik alkalmazási lehetősége, a már említett, és több tudományos konferencián bemutatott kollokációs szótár kidolgozása is.

12. Zusammenfassung

Verwendungsmöglichkeiten der automatischen morphologischen Analyse der kroatischen Standardsprache auf dem deutschen Sprachgebiet

1. Einleitung

Wie schon der Titel verrät, beschäftigt sich diese Arbeit mit einer Sprachvariante. Aufgrund der morphologischen Analyse als Ausgangspunkt umfasst sie aber zur gleichen Zeit zwei, aus verschiedenen Familien stammende Sprachen. Bevor man sich aber mit dem tatsächlichen Thema auseinandersetzt, sind einige Sachen in Bezug auf die Auswahl des Prozesses der morphologischen Analyse und der im Titel erwähnten Zielgruppe zu klären.

1.1. Die Ansätze für die Auswahl des Prozesses der morphologischen Analyse

Nach der Festlegung des *Gemeinsamen Europäischen Referenzrahmens*[143] ist ein großer Bedarf an Lern- und Lehrmaterialien für die kroatische Sprache entstanden, die in Zusammenhang mit dem Referenzrahmen stehen. Bei der kroatischen Ausgabe des Referenzrahmens ist aber zu bemerken, dass der Schwerpunkt der im Rahmen aufgeführten Sprachkompetenzen auf den traditionellen grammatischen Ebenen liegt (Jelaska 2005a: 22). Diese Tatsache wurde auch durch die vielen, in Kroatien begonnenen oder noch immer laufenden Projekte unterstützt, wie z. B. *Hrvatski kao strani jezik: razvojna gramatika i rječnik* (MZOŠ 130738)[144] [Kroatisch als Fremdsprache: Entwicklung der Grammatik und Lexik] und *Hrvatski kao drugi i strani jezik* (MZOŠ 0130438) [Kroatisch als Zweit- und Fremdsprache]. Als Ergebnisse dieser Projekte sind unterschiedliche Materialien entstanden, z. B. ein morphologisches Wörterbuch, ein

[143] Council of Europe (2001). *Common European Framework of Reference for Languages*: *leraning, teaching, assesment*. Cambridge: Cambrige University Press, im Folgenden: Referenzrahmen
[144] http://www.croatiana.org/croatiana--projekti-hrvatskikaostrani.htm, 4. Februar 2008

Konjugationswegweiser sowie ein für die Bestimmung des kroatischen Grundwortschatzes aus 700 häufigsten Wörtern bestehendes Korpus[145]. Diese weisen aber auf das eigentliche Hauptziel solcher und ähnlicher Initiativen hin und das wäre die Erstellung verschiedener Materialien, die zur Entwicklung der Kompetenzen von Lernenden des Kroatischen als Zweit- oder Fremdsprache[146] beitragen würden. Ein zusätzlicher Beweis ist auch die Meinung der Kroatischlehrer, die behaupten, dass „die existierenden Grammatikbücher nicht nur den Bedürfnissen der Lehrer, sondern auch denen der Schüler entsprechen[147]“ (Cvikić 2005a: 320). Es wurde mehrmals betont, dass ein großer Bedarf an Materialien im Sinne von Grammatik- und Übungsbüchern, Lehr- und Sachbüchern, Wörterbüchern und Lektürebüchern besteht (Cvikić 2005a: 320). Daher hat sich als logische Fortsetzung dieser Idee auch im Rahmen dieser Arbeit die Erstellung bzw. Der Erstellungsplan solcher, auf dem Prozess der automatischen morphologischen Analyse beruhenden Materialen erwiesen, die das Lernen und Lehren des Kroatischen unterstützen.

Diese Arbeit erzielt zwei Verwendungsmöglichkeiten der morphologischen Analyse. Die erste ist nur in theoretischem Sinne zu verstehen, nämlich im Konkretisieren der Flexionsparadigmen der kroatischen flektierbaren Wortklassen durch ein morphologisches Wörterbuch, was als Nebenprodukt des formalistischen Vorgehens bei der morphologischen Analyse des Kroatischen entstehen würde. Im Unterschied zu den bereits existierenden Wörterbüchern der kroatischen Sprache würde das computergestützte morphologische Wörterbuch des Kroatischen alle Lemmata aus Vladimir Anić' *Veliki rječnik hrvatskoga jezika*[148] [Großwörterbuch der kroatischen Sprache] enthalten und nicht nur auf den in Grammatikbüchern angeführten Paradigmenbeschreibungen, sondern auch auf den Ergebnissen der Korpusanalyse beruhen. Dadurch könnte es den tatsächlichen Sprachgebrauch widerspiegeln.

Die zweite Anwendungsmöglichkeit der morphologischen Analyse ist die Erstellung maschineller Übersetzungssysteme für die Sprachrichtungen Kroatisch-Deutsch und Kroatisch-Ungarisch[149] (mehr dazu im Kapitel 9). Da das Hauptziel dieser Arbeit die Präsentation jener linguistischen Problematik und der Dilemmas ist, die bei der Entstehung des morphlogischen Parsers aufgetaucht sind, wird die Vorstellung der Verwendungsmöglichkeiten nur auf einer theoretischen Basis besprochen.

[145] Die genauen Daten und Titel sind in der Arbeit zu finden
[146] Falls nicht anders angedeutet, wird im Folgenden statt diesem Begriff nur „Kroatisch“ verwendet.
[147] Übersetzung MA: Originaltext: „Postojeće gramatike za izvorne govornike ne odgovaraju ni potrebama predavača ni potrebama učenika.“
[148] Anić, V. (2000). *Veliki rječnik hrvatskoga jezika*. Zagreb:Novi Liber, im Folgenden RHJ
[149] Der Bedarf an der Anfertigung ungarischer Lehr- und Lernmaterialien beruht auf Barić' (2006) Werk.

1.2. Warum fiel die Wahl auf das deutsche Sprachgebiet?

Da man für die oben genannten, auf morphologischer Analyse basierenden Tools eine Zielgruppe bestimmen musste, wurde die Anzahl der Kroaten im Ausland näher untersucht (Prilog 1 -, Beilage 1‘). Wie aus der Tabelle ersichtlich ist, ist die Anzahl der Kroaten im Ausland auf dem englischen und deutschen Sprachgebiet am höchsten. Kroatisch wird im deutschen Sprachraum auch sehr oft als Zweit- oder sogar Fremdsprache unterrichtet (Jelaska 2005c). Berücksichtigt man 1.) die Tatsache, dass das Auswandern der Kroaten in den deutschen Sprachraum 1918 begann[150], 2.) die Hypothese, dass diejenigen Lernenden, deren Muttersprache nicht kroatisch ist, die größten Probleme mit der Einprägung der kroatischen grammatischen Regeln haben[151] und 3.) dass ähnliche Materialien für die englischsprachigen Lernenden bereits existieren, hat sich logischerweise die Entwicklung der oben erwähnten Tools für die Sprachrichtung Deutsch-Kroatisch-Deutsch ergeben.

1.3. Die Gründe für die Auswahl der morphologischen Analyse als Ausgangsprozess

Aus zwei Gründen wurde die morphologische Analyse als Ausgangsprozess für die weitere Entwicklung der oben genannten Tools gewählt. Zum einen musste man solch einen Prozess wählen, dessen Basis man leicht umwandeln und in viele Systeme implementieren kann. Die Entscheidung fiel auf den unifikationsbasierenden Parser HUMOR[152]. Zum zweiten werden aufgrund des Prozesses der morphologischen Analyse solche Tools entwickelt, die auf der Korpusanalyse und dem tatsächlichen Sprachgebrauch beruhen, was bei den auf dem Prozess der Synthese beruhenden Tools nicht immer der Fall ist.

[150] http://www.mvpei.hr/hmiu/tekst.asp?q=osi001, 4. Februar 2008.
[151] Diese Hypothese beruht auf einer soziolinguistischen Umfrage, die von den Kroatischlehrern im Schulungs- und Kulturzentrum der Ungarn in Kroatien, Osijek im Jahre 2004 durchgeführt wurde. Die Untersuchung hat ergeben, dass die Muttersprache 85% der Schüler Ungarisch ist, und dass nur 24% von ihnen Kroatisch in der alltäglichen Kommunikation verwenden (Aleksa 2007a). Nach der persönlichen Einschätzung der Lehrer hat die Mehrheit der Schüler eine kroatische Kompetenzstufe auf dem Niveau B1 und B2. Für die genauere Bestimmung der Stufen sind weitere Untersuchungen nötig.
[152] High-speed Unification Based Morphology ist ein morphologischer Parser, entwickelt von der Firma MorphoLogic.

2. Analyse der kroatischen Lehr- und Lernmaterialien

Wie schon erwähnt ist das Nebenziel dieser Arbeit die Beschreibung der notwendigen Schritte, die bei der Erstellung des morphologischen Wörterbuches durchzuführen sind. Aus diesem Grund wurde eine Analyse der kroatischen Lehr- und Lernmaterialien durchgeführt, mit dem Hauptziel, die Lexik näher zu untersuchen. Durch die Analyse wurde außerdem die Klarheit der grammatischen Erklärungen, ihre Eindeutigkeit und Gebräuchlichkeit vom Standpunkt der nicht-kroatischen Muttersprachler[153] untersucht.

2.1. Prinzipien der Analyse

Die Analyse wurde in zwei Schritten durchgeführt. Im ersten wurden die kroatischen Lehrwerke und im zweiten die im Kroatischunterricht benutzten Grammatikbücher und sekundäre Literatur untersucht. Die Auswahl der zu untersuchenden Materialien beruht auf den im *Kurikulum hrvatske nastave u inozemstvu*[154] [Das Curriculum des Kroatischunterrichts im Ausland] begründeten Ansätzen.
Dem Ziel dieser Arbeit entsprechend versucht die Lehrwerkanalyse folgende Fragen zu beantworten:

1. Welche Wörter und Wortklassen sind in den analysierten Lehrwerken am häufigsten vertreten, und welchen Stellenwert nehmen sie im kroatischen Häufigkeitswörterbuch ein?
2. Welche grammatischen und lexikalischen Kollokationen (vgl. Kapitel 2.1.3.) kommen in den Lehrwerken am häufigsten vor, und wie sind diese im tatsächlichen Sprachgebrauch vertreten (vgl. Kapitel 6.3.)?

Die Analyse der Grammatikbücher und der sekundären Literatur versucht auf die folgenden Fragen Antworten zu geben:

[153] Unter dem Begriff Schüler oder Lernende werden in dieser Arbeit diejenigen Lernenden verstanden, die über eine sprachliche Kompetenz auf der Mittelstufe oder einer höheren Stufe verfügen.
[154] Das *Kurikulum hrvatske nastave u inozemstvu* ist auf der Internetseite des kroatischen Bildungsministerium zugänglich: http://public.mzos.hr/Default.aspx?sec=2116, 6. Juni 2008, im Folgenden: Kurikulum

3. Tragen die Paradigmen- und morphologischen Beschreibungen sowie die Veranschaulichung der Paradigmen von den in Lehrwerken am häufigsten vertretenen Wörtern und Kollokationsbestandteilen zur erfolgreichen Aneignung der kroatischen Sprache bei? Die Frage wird durch folgende Kriterien beantwortet:
    c) die Eindeutigkeit der Beschreibungen und Veranschaulichungen
    d) die Anwendbarkeit der Beschreibungen und Veranschaulichungen
4. Bieten die Grammatikbücher konkrete Informationen zu den Paradigmen der in den Lehrwerken meistvertretenen Wörter?

Die aufgeworfenen Fragen werden durch die Korpusanalyse beantwortet.

2.1.1. Korpusanalyse

Für die Ausrechnung der Vorkommenshäufigkeit der Wörter wurde das Program *dtsearch*[155] benutzt, während die Kollokationsuntersuchung mit dem *Ngram Statistics Package*[156] durchgeführt wurde. Im Folgenden werden die genannten Programme kurz vorgestellt.

2.1.2. Dtsearch

Dtsearch hat sich as wirkungsvollstes Programm für die Ausrechnung der Vorkommenshäufigkeit der Wörter erwiesen. Durch die schnelle Indexierung des Korpus bietet das Programm verschiedene statistische Ausrechnungen. Zu den Mängeln gehört aber die Unmöglichkeit der Lemmatisierung des kroatischen Korpus, was mithilfe eines Lemmatisierers, den ich in Zusammenarbeit mit MophoLogic entwickelt habe, gelöst wurde.

[155] http://www.dtsearch.com/customDevelop.html, 29. April 2008
[156] http://www.d.umn.edu/~tpederse/nsp.html, 29. April 2008

2.1.3. NSP

Das NSP, oder „Ngram Statistics Package“ ist ein frei erhältliches Computerprogramm auf der Basis der Computersprache Perl, das im Rahmen eines, von Ted Pedersen geleiteten Forschungsprojektes entwickelt wurde. Das Programm selbst ist das Ergebnis der Zusammenarbeit von Mathematikern und Linguisten und beruht auf den Prinzipien der mathematischen und statistischen Analyse. Seine linguistischen Anwendungsbereiche sind sehr breit, was auch viele sprachwissenschaftliche Forschungsmöglichkeiten bietet. Im primären Bereich dient NSP der raschen Untersuchung einer großen Anzahl verschiedener Korpora im Sinne der Suche nach Kollokationen[157], d. h. nach Relationen zwischen zwei, drei oder theoretisch einer unbegrenzten Anzahl von Wörtern oder Wortpaaren, sogar dann, wenn diese durch ein oder mehrere Wörter, bzw. Zeichen getrennt sind. Wichtig ist auch, dass die zu untersuchenden Einheiten oder Tokens nicht nur aus einem Wort bestehen müssen, sondern auch mehrere Wörter enthalten können. Auf diese Weise können auch Relationen zwischen zwei Wörtern (Bigrams), drei (Trigrams) und theoretisch einer unbegrenzten Anzahl von Wörtern (N-grams), aber auch Relationen zwischen einer Wortgruppe als einer zu untersuchenden Einheit oder Token zu anderen Wörtern oder Wortgruppen untersucht werden. Außer den vielen erwähnten Vorteilen hat NSP auch einen Nachteil, nämlich die Sprachbarriere. Das Programm ist für die im morphologischen Sinne flexionsarme englische Sprache entwickelt worden. Daher hat sich die Bearbeitung einer flektierenden Sprache wie des Kroatischen, aber auch der Texte, die in deutscher Sprache verfasst wurden, als schwierig erwiesen. Bevor man solche Texte mit NSP untersucht, müssen sie vorbereitet und bearbeitet bzw. lemmatisiert werden. Für die Lemmatisierung des kroatischen Korpus habe ich in Zusammenarbeit mit MorphoLogic einen kroatischen Analysator entwickelt (vgl. Aleksa 2006).

[157]In dieser Arbeit werden Kollokationen als Wortverbindungen zweier oder mehrerer lexikalischen Einheiten betrachtet, wobei die Wahrscheinlichkeit, dass deren Bestandteile im tatsächlichen Sprachgebrauch immer nebeneinander auftreten, sehr groß ist. Diese Definition beruht auf der im Lemnitzer-Zinsmeister (2006:15) verfassten Definition, wo unter dem Begriff *Kollokation* eigentlich *Kookkurenz* und *Kovorkommen* zu verstehen ist, sowie auf Bensons Bestimmung, dass Kollokationen „arbitrary and recurrent word combitations“ sind (1990, opus.cit.
http://www.latl.unige.ch/personal/vseretan/publ/EURALEX2004_VS_LN_EW.pdf, 20. Januar 2008).

2.2. Analyse der auf dem deutschen Sprachgebiet gebrauchten kroatischen Lehrwerke

Das Ziel dieser Analyse ist anhand des im Kapitel 2.1. Besprochenen und mithilfe der Korpusanalyse Auskunft über den Gebrauch der Lexik in deutschen Lehrwerken zu geben. Die Ergebnisse werden dann mit den Prinzipien der Auswahl der Wörter für die Zusammenstellung eines kroatischen Grundwortschatzes sowie mit grammatischen Informationen, die den Schülern zur Verfügung stehen, verglichen. Das Korpus besteht aus zwei auf deutschem Sprachgebiet gebräuchlichen kroatischen Lehrwerken, die bisher von keinerlei quantitativen und qualitativen Untersuchungen erfasst wurden (Blagus 2005a és 2005b, Cvikić 2005b:221), nämlich das Lehrbuch von S. Drilo, *Kroatisch, TL 1, Lehrbuch für Anfänger* (2006)[158] und *Kroatisch Lehrbuch für Fortgeschrittene* (1996)[159].

Die Untersuchung hat Folgendes ergeben: KA und KF beinhalten 7591 verschiedene Wörter, die auf 3833 Lemmata zurückzuführen sind. Die Abbildungen 1 und 2 veranschaulichen die Aufteilung der Lemmata in den jeweiligen Büchern. Auf Abbildung 3 sind die fragwürdigen Lemmata dargestellt, d. h. diejenigen, die heute nicht mehr oder nicht nur dem kroatischen Lexikon angehören, sondern mehr als Teile des serbischen fungieren (mehr dazu im Kapitel 6.2.).

Die in Lehrwerken benutzte Lexik steht in Zusammenhang mit den Daten aus dem Häufigkeitswörterbuch des Kroatischen. Wenn man bei der Zusammenstellung des morphologischen Wörterbuches oder Parsers diesen Wörtern Vorrang gäbe, würde das Textverständnis oder die –Bearbeitung enorm steigen. Die Ergebnisse des *T-tests* (Prilog 3 - ‚Beilage 3‘) zeigen, dass als grammatische Kollokationsbestandteile am meisten Verben und danach Substantive auftreten, was auf den Bedarf an Paradigmenbeschreibungen dieser zwei Wortklassen verweist. Nach der *Pointwise Mutual Information* Analyse (Prilog 4 - Beilage 4) darf man die adjektivischen Kollokationen nicht außer Acht lassen, da sie im tatsächlichen Sprachgebrauch häufig vorkommen. Mehr Informationen zu den syntaktischen Verhältnissen der Wörter sind aus der tri- und 4-Gram Analyse herauszulesen (Prilog 5 u 6 - ‚Beilagen 5 und 6‘). Die Fragen der Syntax bleiben aber außerhalb des Rahmens dieser Arbeit und werden daher nicht im Detail besprochen.

[158] Im Folgenden: KA
[159] Im Folgenden: KF

2.3. Analyse der im Kroatischunterricht benutzten Grammatikbücher

Die Auswahl der für diese Analyse gewählten Bücher beruht auf den Vorschriften des kroatischen Bildungsministeriums[160] und Aussagen der Kroatischlehrer, die Kroatisch als Zweit- oder Fremdsprache unterrichten[161]. Unter Verwendung des Prinzips der Repräsentativität wurden folgende Werke ausgewählt:

- J. Silić–I. Pranjković (2005) *Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*[162],
- E. Barić et al. (1995) *Hrvatska gramatika*[163],
- D. Raguž (1997) *Praktična hrvatska gramatika*[164],
- die elektronische Version des *Gramatički tezaurus hrvatskoga jezika v 1.2.*[165],
- die Internetversion des kroatischen morphologischen Lexikons *Hrvatski morfološki leksikon*[166].

Es werden nur die Ergebnisse der Untersuchung der flektierbaren Wortklassen präsentiert.

2.3.1. Verben

Nach der Untersuchung der Veranschaulichungen des verbalen Paradigmas und den damit verbundenen Erklärungen ist zu schließen, dass für die Kroatischlernenden eine neue Vorgehensweise nötig ist, die auf der Korpusanalyse beruht und in Kapitel 7.4.2. näher beschrieben wird.

[160] Die ausgewählten Bücher sind im *Kurikulum hrvatske nastave u inozemstvu* aufgelistet, unter § 8.1.3 (http://public.mzos.hr/Default.aspx?sec=2116-a, 8. Juni 2008). Da die anderen angeführten Werke im Kroatischunterricht auf den Stufen B1 und B2 nicht benutzt werden, wurden sie aus der Analyse ausgelassen.
[161] Die Spezifika dieser Arbeit in Betracht genommen, und aufgrund der Tatsache, dass die Kroatischlehrer auf dem deutschen Sprachgebiet nicht erreichbar waren, wurde eine interne Befragung der Kroatischlehrer, die ungarische Muttersprachler Kroatisch unterrichten, durchgeführt.
[162] Im Folgenden: GHJ
[163] Im Folgenden: HG
[164] Im Folgenden: PHG
[165] Im Folgenden: GT
[166] Im Folgenden: HML

2.3.2. Adjektive

Ähnlich wie beim verbalen Paradigma ist es bei der Untersuchung des adjektivischen Paradigmas zum oben angeführten Schluss gekommen. Die wichtigste Frage, die dabei aufkam, betrifft die tatsächliche Anzahl der adjektivischen Lemmata im RHJ, da einige Inkonsequenzen bemerkt wurden (Aleksa 2007b). Eine neue Vorgehensweise ist in Kapitel 7.3. zu sehen.

2.3.3. Substantive

Wenn man das substantivische Paradigma in Grammatik- und Hilfsbüchern sowie die zu den einzelnen Substantiven dargebotenen Informationen aus dem RHJ betrachtet, bemerkt man zuerst den Mangel an notwendigen Informationen und dann die Problematik des grammatischen Genus' (mehr dazu in Kapitel 8.4.2.).

2.3.4. Pronomina und Numeralia

Betrachtet man die in Grammatik- und Hilfsbüchern angeführten Informationen zum pronominalen Paradigma sowie die Beschreibung der Numeralia, lässt sich daraus schließen, dass es an konkreten Informationen für die Lernenden fehlt. Es ist eine neue, durch den HUMOR'schen Formalismus bedingte Veranschaulichung nötig, die auf alle mit Morphologie verbundenen Fragen konkrete Antworten anzubieten fähig ist.

2.3.5. Die Untersuchung Vladimir Anić' kroatischen Großwörterbuches *Veliki rječnik hrvatskoga jezika*

Nach der Untersuchung des kroatischen Großwörterbuches kam ich zum Ergebnis, dass das Wörterbuch nicht genügend Informationen für die Kroatischlernenden bietet, als dass diese daraufhin korrekte Paradigmen aufbauen könnten. Nötig ist ein neues Schülerwörterbuch, das den Schülern nach Prószékys (1997b) Ansätzen auf dynamische Weise vielerlei notwendige Informationen anbietet. Die Grundzüge eines solchen

Wörterbuches beinhaltet auch das im Rahmen dieser Arbeit konzipierte morphologische Computerwörterbuch.

3. Morphologische Analyse im Dienste der Anfertigung des morphologischen Wörterbuches

3.1. Die Grundprinzipien der Entwicklung des morphologischen Computerwörterbuches der kroatischen Standardsprache

Bevor mit der Entwicklung des morphologischen Wörterbuches begonnen wurde, mussten einige Fragen geklärt werden. In diesem Kapitel wird die Auswahl des Verfahrens und der Sprachvariante erklärt.

Da das morphologische Wörterbuch eine Anwendungsmöglichkeit der morphologischen Analyse bedeutet, wäre es wichtig, das Verfahren zu erklären, nämlich die Benutzung des Prozesses der morphologischen Analyse und nicht der Synthese. Anhand der Lehrwerkanalyse, wie schon erwähnt, entstand die Notwendigkeit der Entwicklung solcher Materialien, die im Gegensatz zu den auf der Synthese basierenden, auf dem tatsächlichen Sprachgebrauch beruhen. Tadić (1994: 31) behauptet in seiner Arbeit, dass mit dem Prozess der Synthese nur die Hypothesen der Forscher zu testen sind und das Ziel nicht das tatsächliche Erkennen der Wortformen ist. Da das endgültige Ziel des morphologischen Parsens die 100%ige Erkennung der Texte ist, enthält die lexikalische Basis des Parsers nicht nur die Lemmata aus dem RHJ, sondern er zieht auch die Ergebnisse der Testkorpusanalyse in Betracht (vgl. Kapitel 4.2.). Die zweite Frage, die zu beantworten ist, ist die Frage der Auswahl der Sprachvariante.

3.2 Die Auswahl der Sprachvariante bei der Zusammenstellung des morphologischen Wörterbuches

Die Frage, die bei der Auswahl der passenden Sprachvariante aufkommt, bezieht sich in erster Linie darauf, ob die lexikalische Basis auf die Standardsprache selbst beschränkt werden soll oder ob die Lemmata, die charakteristisch für andere Varietäten des Kroatischen sind, miteinbezogen werden sollen. Die Lehrwerkanalyse hat nämlich ergeben, dass in kroatischen Lehrwerken eine bedeutende Anzahl von Lemmata vertreten

sind, die nicht für die Standardsprache charakteristisch sind (vgl. Kapitel 2.2). Die Wahl fiel auf die standardsprachige Basis, mit der Möglichkeit ihrer Verbreitung durch die für andere Varianten des Kroatischen charakteristischen Lemmata, wenn ihre Notwendigkeit durch die Korpusanalyse bewiesen ist und damit der Texterkennungsgrad im Wesentlichen steigt.

4. HUMOR als eine Art morphologischen Analysators

HUMOR, oder High-Speed Unification based Morphology ist ein von MorphoLogic entwickelter unifikationsbasierter morphologischer Parser, der in erster Linie der morphologischen Analyse von Sprachen dient und als Grundlage vieler maschineller Übersetzungsprogramme fungiert (Prószéky – Kis 1999a). Das Programm wird unter anderem als Basis für Übersetzungssysteme wie MobiMouse, MobiDic und MetaMorpho gebraucht (vgl. Kapitel 9). Bisher wurde HUMOR sowohl für das System der agglutinierenden Sprachen, als auch der flektierenden Sprachen implementiert, was auch die unterschiedlichen sprachlichen Versionen von oben genannten Übersetzungsprogrammen und Übersetzungstools erklären können. Seine Vorteile liegen u. a. in der raschen Durchführung von Aufgaben, aber auch in der Möglichkeit, den Parser allen Sprachsystemen anzupassen. Die erste Demoversion des morphologischen Parsers HUMOR wurde von MorphoLogic im Jahre 1992 entworfen. Das Hauptziel lag nicht in der Entwicklung industrieller Orthographieprüfer, Wortrennungsprogramme und Thesauri, da solche Programme schon seit Jahren auf dem Arbeitsmarkt vertreten sind, sondern in erster Linie im linguistischen Parsen von Sprachen für verschiedene Suchzwecke und das flache bzw. volle Parsen in übersetzungsunterstützenden Systemen (Prószéky–Kis 1999a: 266).

4.1. Die Funktionsweise von HUMOR

Das Hauptprinzip des morphologischen Parsers HUMOR ist die Aufteilung der Lemmata in STEMS und TERMS. Traditionelle morphologische Kategorien wie Wortstamm und Affixe werden hier absichtlich nicht benutzt, da die im Parser definierten Begriffe mit ihnen nicht immer übereinstimmen. Allgemein gesagt umfasst die Kategorie Stem jene

Wortteile, die im Laufe der Flexion unverändert bleiben, während der veränderliche Rest des Lemmas als Term bezeichnet werden kann. Wichtig ist auch die Tatsache, dass ähnlich wie bei der traditionellen Morphologie ein Wort nicht aus einem Ø–Stem, aber aus einem Ø–Term bestehen kann. Am Beispiel des kroatischen Wortes *brzina* - ‚die Geschwindigkeit' - kann man sehen, dass im Gegensatz zu den traditionellen Kategorien *brzin-* als Stem bezeichnet wird, gefolgt von dem Term *–a*, weil der Stem-Wortteil während der Flexion unverändert bleibt. Die detaillierte Beschreibung von HUMOR sowie die ungarischen und kroatischen Spezifika sind in diesem Kapitel der Arbeit gegeben.

4.2. Struktur des morphologischen Parsers

Außer der *Engine*, die als Triebwerk des Programms bezeichnet werden kann, spielt beim Ablauf des Programms auch die zweiteilige lexikalische Basis im Rahmen einer Datenbank eine wichtige und entscheidende Rolle. Die im Lexikon kompilierte Basis der kroatischen Sprache enthält ca. 60 000 lexikalische Einheiten, die aus dem RHJ stammen. Es muss betont werden, dass die Anzahl der lexikalischen Einheiten nicht begrenzt ist, d. h. neue Lemmata können dem Lexikon jederzeit beigefügt bzw. dupliziert werden.

4.3 Begriffsdefinitionen

Bevor ich mit der detaillierten. Beschreibung des Prozesses beginne, sind einige Begriffe zu klären.

1. Weil beim morphologischen Parsen das Leerzeichen die Analysebegrenzung bedeutet und das Wort auf einer graphischen Ebene verstanden wird, entstehen einige Probleme mit der Erkennung aller Phrasenelemente bei zusammengesetzten Phrasen. Die Phrasenbestandteile werden in der ersten Phase nur dann erkannt, wenn sie in die lexikalische Basis eingetragen sind. Z. B. das Verb *kupati* kann mit einem Reflexivpronomen *sebe/se* in der verbalen Phrase *kupati se* ‚sich baden' stehen, oder das Verb kann alleine in der Bedeutung *kupati* ‚baden' registriert werden. Wenn beide Varianten in das Lexikon eingetragen sind, wird das Pronomen bei der Analyse als dem Verb zugehörend erkannt. Da diese Arbeit die

morphologische Beschreibung der kroatischen Sprache zum Ziel hat, werden der syntaktische, phonetische und pragmatische Aspekt beiseite gelegt. Die Erkennung komplexer syntaktischer Strukturen wird durch die Benutzung des mit syntaktischen Kenntnissen bereicherten Parsers HumorEsk effektiver.

2. Im Unterschied zur traditionellen Grammatik wird die Angehörigkeit eines Wortes zu einer Wortklasse durch ihr Paradigma bestimmt (vgl. Kapitel 7).
3. Die Nützlichkeit der im Rahmen des HUMOR ausgearbeiteten formalistischen Lösungen wird durch eine bevorstehende Pilotstudie untersucht. Die Anfertigung des morphologischen Wörterbuches befindet sich in der letzten Phase (vgl. Kapitel 8).
4. Die kroatischen Digraphen *dž*, *lj* und *nj* werden jeweils als Einzelbuchstabe behandelt.

5. Übersicht über weitere morphologische Parser für agglutinierende, flektierende und hochflektierende Sprachen

Die Autoren von HUMOR (Prószéky–Kis 1999a: 261) haben mehrmals die Vorteile der Zwei-Ebenen-Morphologie (Koskenniemi 1983) betont, was auch in der Beschreibung von GERTWOL zu finden ist (Haapalainen – Majorin (1994). Dieses Kapitel erklärt einerseits die Besonderheiten der deutschen Sprache und die von GERTWOL benutzten Lösungen, andererseits vergleicht es diese mit der Problematik der kroatischen und polnischen Sprache. Da eines der zukünftigen Projekte unter anderem die Entwicklung von Übersetzungstools für die ungarische Sprache ist, wird in diesem Kapitel auch die Besonderheit der Bearbeitung agglutinierender Sprachen näher unersucht.

6. Die Anfertigung des kroatischen Analysators

6.1. Die Frage der Sprachgeschichte und Sprachpolitik

Die kroatische Sprache kann, diachronisch gesehen, grob in drei Perioden aufgeteilt werden: die Zeit bis zum 19. Jahrhundert mit der nichtunifizierten Graphologie, die Periode des Serbokroatischen und schließlich die Zeit ab 1990 - die Zeit der heutigen, modernen kroatischen Sprache. Unter diesen drei Sprachvarianten bestehen Unterschiede

auf fast allen Sprachebenen, was zur Problematik der Zusammenstellung einer kroatischen lexikalischen Basis führt. Wenn beschlossen worden wäre, eine morphologische Analyse älterer Sprachvarianten zu ermöglichen (z.B. des Serbokroatischen), müsste man beispielsweise auch die Lexeme aus der serbischen Sprache in das Lexikon inkorporieren. Die Frage ist, ob dieses Lexikon dann noch als kroatisches Lexikon gelten würde. Wenn man anderenfalls für eine lexikalische Basis nur die Lemmata aus dem RHJ nehmen würde, ergäben sich Schwierigkeiten bei der Korpusanalyse und das Programm wäre nicht in der Lage, die vor 1990 entstandenen, in serbokroatischer Sprache verfassten Werke zu analysieren. Das Hauptproblem liegt darin, dass auch heute serbische Lexeme einen Bestandteil des kroatischen Lexikons ausmachen und dass diese in großem Maße auch auf kroatischen Internetseiten vertreten sind. Das andere Problem liegt in der Erkennung älterer kroatischer Varianten. Da die einzige existierende Variante des slawischen HUMOR, die Polnische (Wołosz 2005), u. a. die Verarbeitung von Texten aus dem 18. Jahrhundert ermöglicht, entstand ein solcher Bedarf auch bei der kroatischen Sprache. Wenn man aber die aus dieser Zeit benutzten Grapheme und die Sprachsituation betrachtet, kann man feststellen, dass eine solche Entscheidung meistens zu komplizierten, teils auch unmöglichen Ergebnissen in Bezug auf die Anzahl und Entstehung der Stems und Terms führen würde. Die endgültige Wahl fiel auf die Standardvariante des Kroatischen mit der Möglichkeit der Erweiterung der lexikalischen Basis mit Lemmata, die dann die Texterkennung verbessern würden.

6.2. Die Frage der Sprachvariante

Den Beschlüssen aus Kapitel 3.2. ähnlich. ist bei der Erarbeitung des morphologischen Analysators die Frage der Sprachvariante aufgetaucht. Sollte man sich auf die geschriebene Variante beschränken, oder andere kroatische Varianten inkorporieren, wie z. B. Lemmata, die charakteristisch für die gesprochene oder dialektale Variante sind. Die Wahl fiel auf die Standardvariante mit der Möglichkeit ihrer Erweiterung durch Lemmata aus anderen Varianten des Kroatischen. falls das nötig sein sollte.

6.3. Das Testkorpus

Da das kroatische Nationalkorpus (HNK) nicht verfügbar war, musste im Rahmen dieser Arbeit der Analysator an einem für diesen Zweck auf der Vorlage des HNK zusammengestellten Korpus getestet werden. Das Testkorpus besteht aus fünf Subkorpora, die Texte aus den folgenden Bereichen enthalten:

1. Literatur (Drama, Prosa, Lyrik)
2. Ephemere Werke und Texte aus den Printmedien
3. Elektronische Medien (Internet, Blogs)
4. Fachbücher
5. Lehrbücher

Weitere Informationen bezüglich der Zusammenstellung des Testkorpus sind in der Arbeit zu lesen.

7. Die Anwendung des morphologischen Analysators im Bereich des Kroatischunterrichts

Aus Kapitel 2.3.5 ist ersichtlich, dass ein größerer Bedarf an der Erstellung verschiedener Materialien für das Lehren und Lernen des Kroatischen besteht. Die Grundzüge des CALL (Computer-Aided Language Learning) (Prószéky 1997b) und Prószékys Bestimmungen berücksichtigend sollte im Falle des Kroatischen mit der Erarbeitung solcher Materialien begonnen werden, die den Schülern auf dynamische Weise möglichst viele Informationen über die Paradigmen verschiedener Wörter anbieten können, und eine Verbindung mit Muttersprache sichern. Diesbezüglich wird in diesem Kapitel ein Entwurf für das anzufertigende morphologische Computerwörterbuch des Kroatischen vorgestellt.

7.1. Die Problematik der Bearbeitung der flektierenden Wortformen und Lösungen

Während der Anfertigung des morphologischen Analysators sind einige Probleme der flektierenden Wortformen aufgetreten. Im Rahmen dieser Arbeit werden nur die

wichtigsten präsentiert bzw. diejenigen, die von den traditionellen Repräsentationen abweichen und dank einer neuen Vorgehensweise Hilfe bei der Aneignung der kroatischen Sprache bedeuten könnten.

7.2. Substantive

7.2.1. Substantivisches Paradigma

Wenn ein Kroatischlernender ein substantivisches Paradigma korrekt ableiten will, benötigt er folgende Informationen:

- Repräsentiert das Substantiv einen belebten oder unbelebten Begriff?
- Welchem Deklinationstyp gehört es an (*-a, -e* oder *–i deklinacija*)[167]?
- die Basisform des Substantivs[168], damit das entsprechende Paradigma korrekt abgeleitet werden kann

Mithilfe der im Rahmen von HUMOR ausgearbeiteten Matrix sind die oben genannten Daten, die außerdem schwieriger herauszufinden sind, nicht nötig (Abbildung 24). Man muss zu jedem Substantiv nur folgende Informationen anführen:

- STEM
- Repräsentiert das Substantiv einen belebten oder unbelebten Begriff?
- Die substantivischen TERM-Formen 11 (N.Sg,), 12 (G. Sg.), 21 (N. Pl.) und 22 (G. Pl.)

Die für die HUMOR'sche Matrix obligatorischen grammatischen Angaben zu den Substantiven *more* ‚Meer' oder *dječak* ‚Junge' sollten dann auf folgende Weise veranschaulicht werden: Die vertikale Linie repräsentiert die Trennung des Stems vom Term:

mor|e sr.r (nž., -e,-a,-a,-a)

[167] Die Bestimmung des Deklinationstyps erfolgt anhand der Genitivform eines Substantivs
[168] Nach HG sind Endungen an die Form zu hängen, die wie oben erklärt bestimmt wurde (Barić et al. 1995: 104). Der kroatische Text lautet: „nastavci se dodaju na osnovu, koja se, u pravilu, dobije ako se u gen. jedn. izostavi nastavak".

dječa|k m.r (ž.,-k,-ka,-ci,-ka)

Die Gebräuchlichkeit der in Abbildung 24 angeführten Matrix wird mit einer zukünftigen Pilotstudie untersucht.

7.2.2. Linguistische Problematik

Das größte Problem beim substantivischen Paradigma ist die Bestimmung der Vokativformen und der tatsächlichen Anzahl der Kasus. Da auf morphologischer Ebene Dativ und Lokativformen immer übereinstimmen, sollten im morphologischen Wörterbuch nur sechs Kasus statt sieben dargestellt werden. Die Problematik des Vokativs sowie die ausführliche Problematik der Kasus sind im Kapitel 8.4.2. beschrieben.

7.3. Adjektive

Nach der traditionellen grammatischen Beschreibung der kroatischen flektierbaren Adjektive unterscheidet man mehrere Typen. Die Grammatiker operieren dabei mit Begriffen wie *opisni pridjevi* ‚beschreibende Adjektive', *odnosni pridjevi* ‚Relativadjektive', *gradivni pridjevi* ‚bildende Adjektive' und *posvojni pridjevi* ‚possessive Adjektive' (vgl. Aleksa 2007b). Außerdem werden die Paradigmen anhand der semantischen Kategorisierung des nachfolgenden Substantivs bestimmt, und zwar mit unterscheidenden Merkmalen wie *određeni pridjev* ‚bestimmtes Adjektiv', *neodređeni pridjev* ‚unbestimmtes Adjektiv', *živo* ‚belebt' und *neživo* ‚unbelebt'. Betrachtet man das traditionelle Paradigma, lässt sich dieses wie in Abbildung 25 veranschaulicht, in einer Tabelle mit 93 Zellen darstellen. Die in HUMOR verwendete Lösung operiert nicht mit semantischen, sondern mit morphologischen Merkmalen. Mithilfe von 16 verschiedenen Zellen (Abbildung 36) kann das ganze Paradigma in der in Abbildung 27 angeführten Matrix zusammengefasst werden. Alle Adjektive aus dem RHJ können in 22 verschiedene Gruppen eingeteilt werden, wie in Abbildung 30 zu sehen ist. Als entscheidend haben sich die Adjektivformen 1 und 5 erwiesen. Eine Ausnahme bilden die von Substantiven gebildeten Adjektive, deren Form mit den in der Matrix angeführten identisch ist, aber deren Paradigma die bestimmte Deklination ausschließt.

7.3.1. Steigerung der Adjektive

Betrachtet man die Steigerung der kroatischen Adjektive, ist zu bemerken, dass im Komparativ nicht nur ein Suffix bzw. ein Term dem Stem hinzugefügt wird, sondern dass meistens die Änderungen im Stammvokal, d. h. im Stem selbst erfolgen. Aus diesem Grund wurden die Komparativformen der Adjektive als separate Lemmata betrachtet und dem Lexikon beigefügt. Sie sind mit ihrem ursprünglichen Lemma verbunden.

7.3.2. Linguistische Problematik

Aus den oben angeführten Beispielen kommt man zum Schluss, dass die in HUMOR präsentierten Lösungen sehr kompakt sind. Das adjektivische Deklinationsparadigma mit 93 Zellen wurde z. B. in HUMOR mithilfe von 21 Zellen und 16 verschiedenen Wortformen veranschaulicht. Das bedeutet, dass die morphologische Analyse der Adjektivformen effizienter durchgeführt und das Erlernen der Adjektivdeklination leichter abgeschlossen werden kann. Die Probleme, die dabei entstanden sind, betreffen die Deklination der „weiblichen“ und die Steigerung der „nicht zu steigernden“ Adjektive, was in diesem Kapitel ausführlicher dargestellt wird. Ein gewisses Problem stellen auch die Rechtschreibregeln dar. Für das Adjektiv ‚ungarisch‘ gibt es laut Babić – Finka – Moguš’ (1995) *Hrvatski pravopis* beispielsweise zwei gleichrangige Erscheinungsformen: *mađarski* und *madžarski* (1995: 277). Solche Lemmata müssen daher dupliziert werden, was die Aufnahme von zwei verschiedenen Stems (in diesem Fall *mađarsk* und *madžarsk*) in das Lexikon bedeutet.

7.3.3. Die Bearbeitung des kroatischen Adjektivparadigmas im Vergleich zu den Lösungen aus anderen Sprachen

Die vorgestellte formalistische Lösung wird mit Beispielen aus anderen Sprachen verglichen. Die Abbildung 31 zeigt ein Beispiel aus dem Russischen und es sind auch einige Lösungen aus dem Polnischen vorhanden.

7.4. Verben

7.4.1. Verbales Paradigma

7.4.2. Verbales Paradigma im Rahmen des morphologischen Analysators

Nach den im Kapitel 2.3.1. präsentierten Untersuchungen kommt man zum Schluss, dass das verbale Paradigma eines der meist vertretenen Themen bei der morphologischen Beschreibung der kroatischen Sprache ist. Die Vielfalt und meistens Widersprüche (vgl. Kapitel 2.3.1.) weisen auf die Notwendigkeit einer neuen Vorgehensweise hin, die den Schülern konkrete Fragen und Lösungen anbieten würde. Es gibt insgesamt 12 413 Verben im RHJ, die nach der traditionellen Grammatik in sieben Gruppen mit Klassen aufgeteilt wurden. Das größte Problem bei der Beschreibung des verbalen Paradigmas bereiten die kroatischen Tempora (*aorist, imperfekt*) sowie die Kategorisierung des verbalen Substantivs (*glagolska imenica*) und des verbalen Adjektivs (*glagoski pridjev*). Die von HUMOR vorgeschlagene Lösung operiert mit 96 Konjugationsgruppen, denen alle Verben aus dem RHJ zugeteilt werden können. Eine ähnliche Veranschaulichung wird auf den Abbildungen 35 und 36 gezeigt.

7.4.3. Linguistische Problematik

Ein zusätzliches Problem, wie schon erwähnt, ist die Tatsache, dass aus der Sicht von HUMOR das Leerzeichen Analysebegrenzungen bedeutet. Diese Problematik wird z. T. durch den geplanten Parser Parser HumorEsk gelöst. Bei dem kroatischen verbalen Paradigma muss man berücksichtigen, dass es sehr viele Verben gibt, die ein unregelmäßiges Paradigma haben. Sie können nur einzeln in das HUMOR'sche Lexikon aufgenommen werden.

8. Das morphologische Computerwörterbuch der kroatischen Sprache

8.1. Die Struktur des Wörterbuches

8.2. Technische Lösungen

8.3. Weitere Anwendungsmöglichkeiten des morphologischen Wörterbuches

Unter Berücksichtigung der Ansprüche der Benutzer müsste das geplante Wörterbuch aus einer Datenbank mit mehreren Elementen bestehen. Es sollte dem Benutzer folgende Anwendungsmöglichkeiten bieten:

1. Der Benutzer könnte das Wort aus einer kroatischen Liste auswählen, oder aus einer fremdsprachigen Liste, die automatisch auf das kroatische Wort hinweisen würde.
2. Der Benutzer könnte zwischen mehreren Ansichten wählen, nämlich einer alphabetischen oder *a tergo* Ansicht.
3. Nach der Auswahl eines Lemmas stehen dem Benutzer folgende Informationen zur Verfügung:
    h) „das Wort wird wie x flektiert".
    i) gehört zu der y- Kategorie
    j) grammatische Informationen (Genus, Numerus, usw.…)
    k) STEM bekommt man indem man z-Anzahl der Zeichen vom Ende des Wortes löscht
    l) das ganze Paradigma
    m) das Wort in einem Kontext
    n) die Worttrennungsweise

Das Wörterbuch würde mit zusätzlichen Informationen versehen, wie z. B. mit Kollokationen, was das Ziel einiger zukünftigen Projekte ist.

8.4. Linguistische Fragen und Lösungen

8.4.1. Die Kasus

Während der Anfertigung des Wörterbuches kamen an einigen Stellen Zweifel auf, die gelöst werden mussten. Die Lösungen aber betreffen die bisherige Tradition und verlangen neue Veranschaulichungen einiger Themen, die durch den ausgearbeiteten Formalismus eventuell eine Hilfe beim Erlernen des Kroatischen darstellen könnten. Der erste Zweifel betrifft die kroatischen Kasus, der zweite das grammatische Genus.

Außer dem Problem von Dativ und Lokativ, bleibt einer der größten Zweifel die Vokativform, was ausführlicher in diesem Kapitel besprochen wird. Die Probleme der kroatischen Kasus beruhen meistens auf mangelnden Informationen in den Grammatikbüchern. Da einige Fragen sogar durch die Korpusanalyse nicht beantwortet werden können (z. B. die Frage der Vokativformen), sind sie durch zukünftige Projekte zu lösen.

8.4.2 Die Genera

Die Bestimmung der Genera der Substantive beruht auf Syntax, was das Grammatikbuch von Barić et al (1995:101) *Hrvatska gramatika* beweist. Da die polnischen Linguisten die Existenz von mehr als drei Genera bewiesen haben (Saloni et al. 2007b: 9), war die Frage berechtigt, ob die Informationen zu den traditionellen Genera den Schülern genug Stoff für die Bildung grammatisch korrekter Sätze geboten haben. Da die Antwort negativ war, wurde eine neue Methode nötig, was durch die Bestimmung der Computergenera, d. h. Subklassen der klassischen Genera, gelöst werden muss. Der Lernende wäre dann fähig, korrekte Sätze zu bilden.
Es wurden neun Genera entdeckt, indem die Substantive in den folgenden Satz eingetragen worden sind:

Iako [*biti*] ovdje [imenica] [odnosna zamjenica] volim, ne vidim [pokazna zamjenica]
Obwohl [*sein*] hier [Substantiv] [Relativpronomen] liebe, sehe ich nicht [Pronomen *jener*]

[*dva*] [imenica].
[*zwei*] [Substantiv].

Die Formen der kroatischen Numeralia und Relativpronomen beweisen die Existenz der eigentlichen Genera, was in Abbildung 44 zu sehen ist.

9. Die weiteren Anwendungsmöglichkeiten der morphologischen Analyse

9.1. Lehr- und Lernmaterialien

9.1.1. Ein- und mehrsprachige Wörterbücher

9.1.2. Tools für das Lernen des Ungarischen

9.2. Übersetzungsunterstützende Systeme

Die unifikationsbasierende Basis von HUMOR bietet viele Möglichkeiten für seine Implementierung in verschiedene Systeme. In diesem Kapitel werden u. a. auch die von MophoLogic, auf der Basis der von HUMOR und HumorEsk ausgearbeiteten Tools präsentiert. Es muss aber gesagt werden, dass die vorgestellten Tools nicht als Werbung angesehen dürfen. Die Auswahl wurde aufgrund der Anzahl und Art von Preisen, die sie bekommen haben, getroffen. Zu den vorgestellten Materialien, Tools und Systemen gehören folgende: MobiDic (*multi-dictionary environment)*, MobiMouse (*context-sensitive instant comprehension tool*), Słownik gramatyczny języka polskiego, Rječnik hrvatskoga jezika na CD-romu, DUDEN Deutsches Universalwörterbuch neu, DigiMorf, Helyësen, Elragoz und Metamorpho (*MetaMorpho Dictionary*). Es werden sowohl Vor-, als auch Nachteile der erwähnten Materialien präsentiert.

10. Fazit und Ausblick

Das Verfassen der vorliegenden Arbeit hat zwei Ergebnisse erzielt. Das eine ist praktischer Natur und bedeutet die Anfertigung von verschiedenen Materialien für den Kroatischunterricht (vor allem für das deutschsprachige Gebiet), während das andere Ergebnis im theoretischem Sinne neue Veranschaulichungen einiger Begriffe bietet. Unter Berücksichtigung des Rahmens dieser Arbeit wurden nur die Bereiche bearbeitet, die aus linguistischer Sicht zu neuen Lösungen führten. Zu meinen zukünftigen Plänen zählt u. a. auch die Erstellung eines Kollokationswörterbuches, wovon auf einigen linguistischen Tagungen bereits die Rede war.

Literatura

Aleksa, Melita (2006). Der kroatische HUMOR: Überlegungen zu einer computergestützten morphologischen Analyse der flektierenden Sprachen. In: *Jezikoslovlje 7.1-2*, Osijek: Filozofski fakultet, 141-152.

Aleksa, Melita (2007a). Die Problematik des bilingualen Erstspracherwerbs bei kroatisch-ungarischen Schülern: Eine Untersuchung der schriftlichen Korpora in kroatischer Sprache In: Balaskó Maria − Szatmári Petra (Hrsg.). *LE 59: Sprach- und Literaturwissenschaftliche Brückenschläge - Vorträge der 13. Jahrestagung der GESUS in Szombathely, 12.-14. Mai 2004 Szombathely*. München: Lincom Verlag, 209-217.

Aleksa Melita (2007b). A horvát melléknévi rendszer nyelvtani leírása és automatikus számítógépes kezelése. In: Heltai Pál (ur.) *MANYE XVI. Nyelvi modernizáció.Vol 3/2.* Pécs-Gödöllő, 1257-1263.

Anić, Vladimir (2000). *Veliki rječnik hrvatskoga jezika*. Zagreb:Novi Liber.

Babić, Stjepan (1990). *Hrvatska jezikoslovna čitanka*, Zagreb: Nakladni zavod Globus.

Babić, Stjepan (1995). *Hrvatski jučer danas sutra*, Zagreb: Školske novine.

Babić, Stjepan (2004). *Hrvanja hrvatskoga*. Zagreb: Školska knjiga.

Babić, Stjepan, Božidar Finka, Milan Moguš (1995). *Hrvatski pravopis*. Zagreb: Školska knjiga.

Bakos Ferenc (2002): *Idegen szavak és kifejezések szótára*. Budapest: Akadémiai kiadó.

Barić, Ernest (2006). *Rode, a jezik?*. Pečuh: Biblioteka Nova.

Barić, Eugenija et al. (1995). *Hrvatska gramatika*. Zagreb: Školska knjiga.

Blagus, Vlatka (2005a). Odabir riječi u udžbenicima hrvatskoga za strance. In: Jelaska, Zrinka. (ur.) *Hrvatski kao drugi i strani jezik*. Zagreb: Hrvatska sveučilišna naklada, 261-273.

Blagus, Vlatka (2005b). Pregled udžbenika hrvatskoga za strance. In: Jelaska, Zrinka. (ur.).*Hrvatski kao drugi i strani jezik*. Zagreb: Hrvatska sveučilišna naklada, 226-234.

Brodnjak, Vladimir (1991). *Riječnik razlika između hrvatskoga i srpskoga jezika*, Zagreb: Školske novine.

Bulaja, Zvonimir (1999). *Klasici hrvatske književnosti I. –epika, romani, novele* CD-rom. Zagreb: Bulaja naklada – ALTF4

Council of Europe (2001). *Common European Framework of Reference for Languages*: *leraning, teaching, assesment*. Cambridge: Cambrige University Press.

Cvikić, Lidija (2005a). Hrvatski kao drugi i strani jezik: stanje i potrebe. In: Jelaska, Z. (ur.) *Hrvatski kao drugi i strani jezik*. Zagreb: Hrvatska sveučilišna naklada, 311-329.

Cvikić, Lidija (2005b). Pregled priručnika hrvatskoga kao drugoga i stranoga jezika. In: Jelaska, Z. (ur.) *Hrvatski kao drugi i strani jezik*. Zagreb: Hrvatska sveučilišna naklada, 219-225.

Danolić, Josip (2000). Raskid s ovostoljetnom tradicijom hrvatskih vukovaca. *Vjesnik,* 6. lipnja, 14.

Drilo, Stjepan (1996). *Kroatisch Tl.2. Lehrbuch für Fortgeshrittene*. Heidelberg: Julius Groos Verlag.

Drilo, Stjepan (2006). *Kroatisch Tl.1. Lehrbuch für Anfänger*. Tübingen: Julius Groos Verlag.

Duden (1997). Deutches Universalwörterbuch A-Z neu. Mannheim: Biblioraphisches Institut & FA Brockhaus.

Farkas Ernő, Mátyás, Naszódi (1990). *Magyar nyelvű mondatok elemzése természetes nyelvű interface céljából*. Budapest: MTA SZTAKI.

Haapalainen, Mariikka, Ari Majorin (1994). *GERTWOL: Ein System zur automatischen Wortformerkennung deutscher Wörter*. Lingsoft, Inc., http://www.lingsoft. fi/cgi-pub/gertwol, 6. kolovoza 2008.

Hašek, Jaroslav (2004). *Doživljaji dobrog vojnika Švejka za svjetskog rata*. Zagreb: Biblioteka jutarnjeg lista.

Hrvatski morfološki leksikon, http://hml.ffzg.hr/hml/, 1. listopada 2008.

Jelaska, Zrinka (2003). Hrvatski glagoli – oblici, Zagreb: Sveučilišna škola hrvatskoga jezika i kulture.

Jelaska, Zrinka (2005a). Jezik – znanje ili sposobnost. In: Jelaska, Z. (ur.) *Hrvatski kao drugi i strani jezik*. Zagreb: Hrvatska sveučilišna naklada, 11-24.

Jelaska, Zrinka (2005b). Oblici hrvatskih riječi. In: Jelaska, Z. (ur.) *Hrvatski kao drugi i strani jezik*. Zagreb: Hrvatska sveučilišna naklada, 136-143.

Jelaska, Zrinka (2005c). *Hrvatski kao drugi i strani jezik*. Zagreb: Hrvatska sveučilišna naklada.

Jelaska, Zrinka, Lidija Cvikić (2003). *Morko – Hrvatski višejezični morfološki rječnik.* Zagreb: Sveučilišna škola hrvatskoga jezika i kulture.

Jelaska, Zrinka, Lidija Cvikić, Jasna Novak-Milić (2005). *Rječnici za učenje hrvatskoga*. In: Jelaska, Z. (ur.) *Hrvatski kao drugi i strani jezik*. Zagreb: Hrvatska sveučilišna naklada, 196-205.

Jonke, Ljudevit (2005). *O hrvatskome jeziku*. Zagreb: Pergamena.

Karttunen, Lauri (1993). Finite-state lexicon compiler. In: *Technical Report* ISTL-NLTT, Palo Alto: Xerox PARC.

Kartunnen, Lauri, Kenneth Beasley (2001). *A short history of two-level morphology*. Proceedings of the ESSLLI 2001, http://www2.parc.com/istl/members/karttune/publications/esslli-2001/twol-history.pdf, 12. rujna 2008.

Kolenić, Ljiljana (2003). *Pogled u strukturu hrvatske gramatike (od Kašićeve do Tkalčevićeve).* Osijek: Sveučilište Josipa Jurja Strossmayera, 2003.

Koskenniemi, Kimo (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. Publication No. 11. Heslinki: University of Helsinki.

Kurikulum hrvatske nastave u inozemstvu: http://public.mzos.hr/Default.aspx?sec=2116, 6. lipnja 2008.

Lemnitzer, Lothar, Heike Zinsmeister (2006). *Korpuslinguistik. Eine Einführung.* Tübingen: Gunter Narr Verlag.

Lohde, Mihael (2006). *Wortbildung des modernen Deutschen.* Tübingen. Günter Narr Verlag.

Lončarić, Mijo (1977). *O čestotnim rječnicima i čestotniku hrvatskog književnog jezika*, Suvremena lingvistika 15-16, Zagreb, 39-48.

Mihaljević, Milica (1993). *Hrvatsko računalno nazivlje*. Hrvatska sveučilišna naklada, Zagreb.

Moguš, Milan (1995).*Povijest hrvatskoga književnoga jezika*. Zagreb: Globus.

Moguš, Milan, Bratanić, Maja, Tadić, Marko (1999). *Hrvatski čestotni rječnik*. Zagreb: Školska knjiga.

Novák Attila (2003). Milyen a jó Humor? In: Alexin Zoltán; Dóra Csendes (szerk.). *Az 1. Magyar Számítógépes Nyelvészeti Konferencia előadásai*, Szeged: SZTE, 138–145.

Novák Attila, Nóra Wensky (2007). Mire jó és hogyan készül egy számítógépes morfológia In: Alberti Gábor, Fóris Ágota (eds.). *A mai magyar formális nyelvtudomány műhelyei*. Budapest: Nemzeti Tankönyvkiadó, 157–169.

Novak-Milić, Jasna (2005). Djelotvornost gramatičkoga poučavanja. In: Jelaska, Z. (ur.).*Hrvatski kao drugi i strani jezik*. Zagreb: Hrvatska sveučilišna naklada, 353-361.

Nowson, Scott, Jon Oberlander, Alistair J. Gill (2005). Weblogs, genres and individual differences. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*,. Stresa, Italy, 1666-1671.

Prószéky Gábor (1994). Industrial Applications of Unification Morphology. *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP).*, Stuttgart: University of Stuttgart, 157–159.

Prószéky Gábor (1997a). Morpho-syntactic Segmentation in Continuous Speech Recognition In: *Proceedings of the 2nd International Workshop on Speech and Computer (SPECOM).*, Cluj-Napoca, 143–146.

Prószéky Gábor (1997b). Language Technology in the Service of CALL. In: Kohn, János; Bernd Rüschoff; Dieter Wolff (eds). *New Horizons in CALL (Proceedings of EUROCALL 96).*, 53–64. Szombathely: Berzsenyi Dániel College

Prószéky Gábor (1997c). Újra papír? Lexikonok, enciklopédiák, szótárak - másképp. In: Polyák Ildikó (szerk.). *VII. Országos Alkalmazott Nyelvészeti Konferencia (MANYE).,* I., Budapest: Külkereskedelmi Főiskola, 23–27.

Prószéky Gábor (1997d). Szótárírási szempontok a számítógépes nyelvi programok korában és korábban. In: Kiss Gábor; Zaicz Gábor (szerk.). *Szavak - nevek - szótárak (Írások Kiss Lajos születésnapjára).,* Budapest: Tinta, 326–335

Prószéky Gábor (1999). Language Technology Tools in the Translator's Practice. *Journal of Computing and Information Technology,* 7(3)., 221–227.

Prószéky Gábor (2001). A nyelvtechnológia és a modern nyelvészet viszonyáról In: Andor J., Szűcs T., Terts I. (szerk.). *Színes eszmék nem alszanak. (Szépe György 70. születésnapjára).* Pécs: Lingua Franca Csoport, 991–998.

Prószéky Gábor (2003). Nyelvi technológiák és gépi fordítás. *Alkalmazott nyelvtudomány* III(1)., 5–11.

Prószéky Gábor, Balász, Kis (1999a). A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly). inflectional languages. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.* Maryland, USA: College Park, 261-268. http://www.morphologic.hu/h_pgpub.htm, 2. kolovoza 2006.

Prószéky Gábor, Balázs, Kis (1999b). Számítógép-hálózat a fordítástámogatásban. *Új Alaplap* XVII/4. 24-27.

Prószéky Gábor; Balázs, Kis (2002). Development of a Context-Sensitive Dictionary. In: Anna Braasch; Alex Povlsen (eds). *Proceedings of the 10th International*

Congress of the European Association for Lexicography (EURALEX)., Vol. I, Copenhagen, 281–290.

Raguž, Dragutin (1997). *Praktična hrvatska gramatika*. Zagreb: Medicinska naklada.

Saloni, Zygmunt (2001). *Czasownik polski*. Warszawa : Wiedza Powszechna

Saloni, Zygmunt, Włodzimierz Gruszczyńsky, Marcin Woliński, Robert Wołosz (2007a). *Słownik gramatyczny języka polskiego*. Warszawa: Wiedza Powszechna

Saloni, Zygmunt, Włodzimierz Gruszczyńsky, Marcin Woliński, Robert Wołosz (2007b). Grammatical Dictionary of Polish. Presentation by the Authors. *Studies in Polish Linguistics*, volume 4, 5-27.

Silberztein, Max (1993). *Dictionnaire electriques et analyse automatique de textes.* Paris: Masson.

Silić, Josip, Branko Ranilović, Salven Batnožić (1996). *Gramatički tezaurus hrvatskoga jezika v1.2*. Zagreb: SYS i Matica Hrvatska, elektroničko izdanje.

Silić, Josip, Ivo Pranjković (2005).*Gramatika hrvatskoga jezika za gimnazije i visoka učilišta*, Zagreb: Školska knjiga.

Skoumalová, Hana (1997). *A Czech Morphological Lexicon*; http://arxiv.org/PS_cache/cmp-lg/pdf/9707/9707020v1.pdf, 18. srpnja 2008.

Steinacker, Ingeborg, Harald Trost (1982). *Parsing German*. In: Horecky, J. (ed).*COLING 82* Academia: North-Holland PUblishing Company, 365-370.

Stojaković, Biljana (2004). Hrvatsko računalno nazivlje u društvenoj interakciji. *Strani jezici* 33/1-2; 83-92.

Šonje, Jure (2000). *Rječnik hrvatskoga jezika*, Zagreb: Grafički zavod Hrvatske.

Tadić, Marko (1994). *Računalna obrada morfologije hrvatskoga književnog jezika*. Ph.D. Thesis, Manuscript. Zagreb: Filozofski fakultet Sveučilišta u Zagrebu. http://www.hnk. ffzg.hr/mt/, 2. kolovoza 2006.

Tadić, Marko (1998). Raspon, opseg i sastav korpusa suvremenoga hrvatskoga jezika, *Filologija* 30-31, 337-347.

Tadić, Marko, Sanja Fulgosi, (2003). Building the Croatian Morphological Lexicon. In: *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages* Budimpešta: ACL, 41-46.

Težak, Stjepko (1991). *Hrvatski naš svagda(š)nji.* Zagreb: Školske novine.

Težak, Stjepko (1995). *Hrvatski naš osebujni*. Zagreb: Školske novine.

Težak, Stjepko (1999). *Hrvatski naš (ne)zaboravljeni*. Zagreb: Tipex.

Težak, Stjepko (2004). *Hrvatski naš (ne)podobni*, Zagreb: Školske novine.

Težak, Stjepko, Stjepan Babić (2005). *Gramatika hrvatskoga jezika: priručnik za osnovno jezično obrazovanje*. Zagreb: Školska knjiga.

Volk, Martin (1999). Choosing the right lemma when analysing German nouns In: Gippert, Jost, Peter Oliveir, (Hrsg.). *GLDV '99 Multilinguale Corpora: Codierung, Strukturierung, Analyse*, Prague: Enigma Corporation, 304-310.

Wołosz, Robert (2005). *Efektywna metoda analizy i syntezy morfologicznej w języku polskim*. Warszawa: Akademicka Oficyna Wydawnicza EXIT.

Znika, Marija (2005). Status glagolskih pridjeva. In: *Rasprave Instituta za hrvatski jezik i jezikoslovlje* 31, 429-440.

Зализняк, Андрей Анатольевич. (1987). *Грамматический словарь русского языка,* М.: Русский язык.

Prilozi

Prilog 1: Brojčano stanje hrvatskih iseljenika i njihovih potomaka u svijetu

1.	Sjedinjene Američke Države	1 000 000-1200 000
2.	Njemačka	350 000-400 000
3.	Australija	250 000
4.	Argentina	200 000-250 000
5.	Kanada	200 000-250 000
6.	Čile	150 000-200 000
7.	Mađarska	90 000
8.	Švicarska	80 000
9.	Austrija	80 000-100 000
10.	Italija	50 000-60 000
11.	Francuska	40 000
12.	Novi Zeland	40 000
13.	Švedska	25 000
14.	Brazil	20 000-30 000
15.	Južnoafrička Republika	8 000
16.	Belgija	6 000
17.	Ekvador	6 000
18.	Nizozemska	6 000
19.	Peru	6 000
20.	Urugvaj	5 000
21.	Velika Britanija	5 000
22.	Venezuela	5 000
23.	Norveška	2 000
24.	Danska	1 000
25.	Luxemburg	1 000
26.	Paragvaj	1 000
27.	Bolivija	500-1000

Izvor: http://www.mvpei.hr/hmiu/tekst.asp?q=osi001, 4. veljače 2008.

Prilog 2: Primjer glagolske paradigme iz GHJ (Silić-Pranjković 2005:46)

GLAGOLSKE VRSTE				
	Inf. osn.	**Prez. osn.**	**Infinitiv**	**Prezent**
Prva vrsta	korijen + suf. -Ø-	korijen + suf. -ē-		1. l. jed. i 3. l. mn
Prvi razred	*-ved-Ø-*	*-ved-ē-// -ved-Ø*	*(dò)vesti*	*dovèdēm/ dovèdū*
	-vez-Ø-	*-vez-ē-// -vez-Ø-*	*(dò)vesti*	*dovèzēm/ dovèzū*
	-plet-Ø-	*-plet-ē-// -plet-Ø-*	*plȅsti*	*plètēm/ plètū*
	-griz-Ø-	*-grīz-ē-// -grīz-Ø-*	*grȉsti*	*grízēm/ grizū*
	-greb-Ø-	*-greb-ē-// -greb-Ø-*	*grȅpsti*	*grèbēm/ grèbū*
Drugi razred	*-žē-Ø-*	*-žm-ē-// -žm-Ø-*	*žêti*	*žmêm/ žmû*
	-čē-Ø-	*-čn-ē-// -čn-Ø-*	*(pò)čēti*	*pòčnēm/ pòčnū*
	-že-Ø-	-žanj-ē-// -žanj-Ø-	žȅti	žänjēm/ *žänjū*
Treći razred	*-lē-Ø-*	*-kūn-ē-// -kūn-Ø-*	*klêti*	*kùnēm/ kùnū*
Četvrti razred	*-sla-Ø-*	*-šalj-ē-// -šalj-Ø-*	*slàti*	*šäljēm/ šäljū*
	-la-Ø-	*-olj-ē-// -olj-Ø-*	*klàti*	*köljēm/ köljū*
	-zva-Ø-	*-zov-ē-// -zov-Ø-*	*zväti*	*zòvēm/ zòvū*
	-ra-Ø-	*-er-ē-// -er-Ø-*	*brȁti*	*bèrēm/ bèrū*
Peti razred	*-r-Ø-*	*-ar-ē-// -ar-Ø-*	*tȓti*	*tärēm/ tärū*
Šesti razred	*-lje-Ø-*	*-elj-ē-//* -elj-Ø-	*mljȅti*	*mèljēm/ mèljū*
Sedmi razred	*-drije-Ø-*	*-dr-ē-// -dr-Ø-*	*(prò)drijeti*	*pròdrēm/ pròdrū*
Osmi razred	*-nije-Ø-*	*-nes-ē-// -nes-Ø-*	*(dò)nijeti*	*donèsēm/ donèsū*
Deveti razred	*-smje-Ø-*	*-smij-ē-/ -smij-Ø-*	*smjȅti*	*smïjēm/ smïjū*
Deseti razred	*-pi-Ø-*	*-pij-ē-// -pij-Ø-*	*pȉti*	*pïjēm/ pïjū*
	-ču-Ø-	*-čuj-ē-// -čuj-Ø-*	*čȕti*	*čüjēm/ čüjū*
Jedanaesti razred	*-mog-Ø-*	*-mož-ē-// -mog-Ø-*	*mȍći*	*mògu/ mȍžēš*
	-rek-Ø-	*-reč-ē-// -rek-Ø-*	*rȅći*	*rèčēm/ rèkū*
	-vrh-Ø-	*-vȓš-ē-// -vȓh-Ø-*	*vȓći*	*vȓšēm/ vȓhū*
Dvanaesti razred	*-sū-Ø-*	*-sp-ē-// -sp-Ø-*	*(pò)sūti*	*pòspēm/ pòspū*
Trinaesti razred	*-sta-Ø-*	*-sta-n-ē-// -stan-n-Ø-*	*stȁti*	*stänēm/ stänū*
	-mog-Ø-	*-mog-n-ē-//-mog-n-Ø-*	*(pò)moći*	*(pò)mognēm/ (pò)mognū*
	-rek-Ø-	*-rek-n-ē-// -rek-n-Ø-*	*(pò)reći*	*pòreknēm/ pòreknū*
Četrnaesti razred	*-da-Ø-*	*-da-d-ē-// -dā-d-Ø-*	*dȁti*	*dádēm/ dádū*

Prilog 3: Rezultati analize bigrama udžbeničkoga korpusa s udžbenicima KA i KF. Prikazano je prvih 100 najčešćih kolokacija nelematiziranoga teksta prema *T-score* analizi.

```
kod<>kuće<>1 8.5047 73 154 89
su<>se<>2 8.2650 100 713 992
u<>seminaru<>3 7.0248 52 1053 52
je<>li<>4 6.9828 77 1908 336
günther<>i<>5 6.9423 55 126 1137
i<>helga<>6 6.9310 54 1137 110
gdje<>je<>7 6.9073 59 127 1908
što<>je<>8 6.8730 70 267 1908
da<>je<>9 6.8488 85 467 1908
ja<>sam<>10 6.4296 44 171 322
i<>h<>11 6.2767 43 1137 66
koliko<>je<>12 6.1977 49 120 1908
jeste<>li<>13 6.1869 39 44 336
g<>i<>14 6.1831 43 88 1137
on<>je<>15 6.0639 49 140 1908
to<>je<>16 5.9153 67 397 1908
jesi<>li<>17 5.5189 31 33 336
u<>zagrebu<>18 5.4970 32 1053 35
da<>se<>19 5.1980 47 467 992
kada<>je<>20 5.1331 40 161 1908
i<>ja<>21 5.1099 35 1137 171
cijeli<>dan<>22 5.0901 26 30 62
u<>kinu<>23 5.0619 27 1053 27
ja<>nisam<>24 5.0077 26 171 111
u<>uredu<>25 4.9572 26 1053 28
jesu<>li<>26 4.9522 25 29 336
ana<>je<>27 4.9159 34 114 1908
književni<>jezik<>28 4.7908 23 26 38
nego<>što<>29 4.7880 24 83 267
vi<>ste<>30 4.6439 22 89 100
ivan<>je<>31 4.5895 32 129 1908
što<>sam<>32 4.5782 25 267 322
s<>kim<>33 4.5591 21 209 21
sam<>se<>34 4.4872 34 322 992
oni<>su<>35 4.4530 23 94 713
ona<>je<>36 4.4117 29 112 1908
koga<>je<>37 4.3544 24 57 1908
mi<>smo<>38 4.3445 20 141 165
su<>razgovarali<>39 4.2922 20 713 46
je<>bilo<>40 4.2821 26 1908 89
ne<>znam<>41 4.1892 18 420 22
kao<>što<>42 4.1846 19 116 267
ne<>mogu<>43 4.1479 18 420 39
tko<>je<>44 4.1434 21 43 1908
čija<>je<>45 4.1012 19 24 1908
u<>zagreb<>46 4.0965 18 1053 24
ovo<>je<>47 4.0932 23 72 1908
koji<>je<>48 4.0920 25 97 1908
kao<>da<>49 4.0540 19 116 467
ti<>si<>50 4.0368 17 124 117
u<>institutu<>51 4.0041 17 1053 19
to<>nije<>52 4.0003 21 397 274
bili<>u<>53 3.9915 19 62 1053
li<>se<>54 3.9845 30 336 992
sam<>rekao<>55 3.9641 17 322 83
o<>čemu<>55 3.9641 16 217 27
```

o<>tome<>56 3.9627 16 217 28
na<>izlet<>57 3.9549 16 409 18
bi<>bilo<>58 3.9525 16 87 89
je<>to<>59 3.9383 45 1908 397
bila<>u<>60 3.9382 19 71 1053
ne<>može<>61 3.8970 16 420 40
rekao<>da<>62 3.8925 17 83 467
bio<>u<>63 3.8830 20 102 1053
gospodin<>braun<>64 3.8673 15 41 22
smo<>se<>65 3.8344 22 165 992
čiji<>je<>66 3.7939 17 29 1908
je<>vrlo<>67 3.7824 22 1908 91
od<>koga<>68 3.7809 15 255 57
u<>menzi<>69 3.7729 15 1053 15
marija<>je<>70 3.7500 20 69 1908
je<>star<>71 3.7372 17 1908 34
kada<>su<>72 3.7129 19 161 713
i<>ti<>73 3.6988 20 1137 124
mislim<>da<>74 3.6927 14 16 467
u<>kojoj<>75 3.6862 15 1053 28
ću<>se<>76 3.6806 18 98 992
studenti<>su<>77 3.6721 16 75 713
je<>već<>78 3.6638 23 1908 116
helga<>su<>79 3.6522 17 111 713
otišao<>je<>80 3.6373 16 31 1908
u<>knjižnici<>81 3.6312 14 1053 16
će<>se<>82 3.6287 20 155 992
bi<>se<>83 3.6096 17 87 992
je<>studentica<>84 3.6009 17 1908 46
prvi<>put<>85 3.5997 13 20 43
na<>fakultetu<>86 3.5666 13 409 14
na<>fakultet<>87 3.5416 13 409 23
je<>bio<>88 3.5408 21 1908 102
bismo<>se<>89 3.5270 14 33 992
svi<>su<>90 3.5037 13 21 713
bio<>kod<>91 3.4987 13 102 154
u<>dubrovniku<>92 3.4981 13 1053 15
i<>mi<>93 3.4567 19 1137 141
jedan<>sat<>94 3.4485 12 96 23
u<>split<>95 3.4408 13 1053 23
si<>to<>96 3.4371 14 117 397
prije<>nego<>97 3.4230 12 70 83
ne<>bih<>98 3.4198 13 420 65
ali<>su<>99 3.4078 15 103 713
bilo<>je<>100 3.4033 19 89 1908

Prilog 4: Rezultati analize bigrama udžbeničkoga korpusa s udžbenicima KA i KF. Pikazano je prvih 100 najčešćih kolokacija nelematiziranoga teksta prema *Pointwise Mutual Information* analizi

```
bolesnoj<>ženi<>1 14.3151 2 2 2
crne<>smrče<>1 14.3151 2 2 2
sedmog<>srpnja<>1 14.3151 2 2 2
osmog<>kolovoza<>1 14.3151 2 2 2
tramvajsku<>stanicu<>1 14.3151 2 2 2
ljetnu<>pozornicu<>1 14.3151 2 2 2
otočić<>noktić<>1 14.3151 2 2 2
devetoga<>jedanaestog<>1 14.3151 2 2 2
kletvom<>krvne<>1 14.3151 2 2 2
jurišinoj<>pjesmi<>1 14.3151 2 2 2
majčica<>ručica<>1 14.3151 2 2 2
knjižica<>majčica<>1 14.3151 2 2 2
četvrtoga<>petog<>1 14.3151 2 2 2
jure<>kaštelan<>1 14.3151 2 2 2
bacati<>baciti<>1 14.3151 2 2 2
šumskih<>bjesova<>1 14.3151 2 2 2
tvojoj<>velikoj<>1 14.3151 2 2 2
momčić<>novčić<>1 14.3151 2 2 2
četvrtog<>travnja<>1 14.3151 2 2 2
njezinim<>učiteljem<>1 14.3151 2 2 2
tridesetoga<>dvanaestog<>1 14.3151 2 2 2
pomogente<>siromašnu<>1 14.3151 2 2 2
državica<>knjižica<>1 14.3151 2 2 2
njihovim<>poznanikom<>1 14.3151 2 2 2
djevojkama<>obama<>1 14.3151 2 2 2
orlić<>uglić<>1 14.3151 2 2 2
kačić<>miošić<>1 14.3151 2 2 2
travčica<>koščica<>1 14.3151 2 2 2
jedanaestog<>studenog<>1 14.3151 2 2 2
bilden<>sie<>1 14.3151 2 2 2
oštri<>zavoji<>1 14.3151 2 2 2
ritter<>vitezović<>1 14.3151 2 2 2
travnja<>četvrte<>1 14.3151 2 2 2
brodić<>poslić<>1 14.3151 2 2 2
nerazumljivim<>željama<>1 14.3151 2 2 2
loša<>vladanja<>1 14.3151 2 2 2
poluglasovi<>odrazili<>1 14.3151 2 2 2
listopada<>pedesete<>1 14.3151 2 2 2
rada<>nezdrava<>1 14.3151 2 2 2
dvanaestog<>prosinca<>1 14.3151 2 2 2
andrija<>kačić<>1 14.3151 2 2 2
šestoga<>sedmog<>1 14.3151 2 2 2
vladanja<>sretne<>1 14.3151 2 2 2
sie<>satze<>1 14.3151 2 2 2
tvog<>kolege<>1 14.3151 2 2 2
velikoj<>brizi<>1 14.3151 2 2 2
slamnigovoj<>poeziji<>1 14.3151 2 2 2
vašoj<>udobnoj<>1 14.3151 2 2 2
kolovoza<>osme<>1 14.3151 2 2 2
novčić<>trbušić<>1 14.3151 2 2 2
sedmoga<>šestog<>1 14.3151 2 2 2
stančić<>prozorčić<>1 14.3151 2 2 2
kupovali<>poklone<>1 14.3151 2 2 2
poslić<>orlić<>1 14.3151 2 2 2
naporna<>rada<>1 14.3151 2 2 2
kućica<>pjesmica<>1 14.3151 2 2 2
```

koščica<>cjevčica<>1 14.3151 2 2 2
gorskom<>kotaru<>1 14.3151 2 2 2
eulerova<>mehanika<>1 14.3151 2 2 2
grančica<>travčica<>1 14.3151 2 2 2
bartol<>kašić<>1 14.3151 2 2 2
srpnja<>sedme<>1 14.3151 2 2 2
1919<>1990<>1 14.3151 2 2 2
vladimir<>nazor<>1 14.3151 2 2 2
pjesmica<>državica<>1 14.3151 2 2 2
uglić<>otočić<>1 14.3151 2 2 2
poeziji<>jurišinoj<>1 14.3151 2 2 2
čvoru<>crne<>1 14.3151 2 2 2
trbušić<>rožić<>1 14.3151 2 2 2
prozorčić<>momčić<>1 14.3151 2 2 2
petog<>svibnja<>1 14.3151 2 2 2
učiteljem<>njihovim<>1 14.3151 2 2 2
kaštelan<>1919<>1 14.3151 2 2 2
dosadašnjim<>radom<>1 14.3151 2 2 2
stolić<>brodić<>1 14.3151 2 2 2
gustav<>matoš<>1 14.3151 2 2 2
krvne<>silnike<>1 14.3151 2 2 2
dokazivati<>dokazati<>1 14.3151 2 2 2
dubrovački<>večernji<>2 13.7301 3 3 3
uskim<>ulicama<>2 13.7301 3 3 3
beskrajan<>niz<>2 13.7301 3 3 3
shvaćamo<>njihove<>2 13.7301 3 3 3
qualitas<>occulta<>2 13.7301 3 3 3
mišljenje<>kritizirali<>2 13.7301 3 3 3
sadašnje<>potrebe<>2 13.7301 3 3 3
objavljenu<>objavljenom<>2 13.7301 3 3 3
visoka<>rasta<>2 13.7301 3 3 3
obavještavaju<>čitatelje<>2 13.7301 3 3 3
podnevna<>pauza<>2 13.7301 3 3 3
zlatan<>barjak<>2 13.7301 2 2 3
gramatici<>jonkeovoj<>2 13.7301 2 3 2
posebnostima<>glasovi<>2 13.7301 2 2 3
posebnom<>dozvolom<>2 13.7301 2 2 3
lakoatletskim<>disciplinama<>2 13.7301 2 2 3
šestog<>lipnja<>2 13.7301 2 2 3
maretićevoj<>gramatici<>2 13.7301 2 2 3
zamišljen<>pričajući<>2 13.7301 2 3 2
naravni<>odrezak<>2 13.7301 2 2 3
siromašnu<>čovjeku<>2 13.7301 2 2 3
sazna<>ljutit<>2 13.7301 2 2 3

Prilog 5: Rezultati analize trigrama udžbeničkoga korpusa s udžbenicima KA i KF. Prikazano je prvih 100 najčešćih kolokacija nelematiziranoga teksta.

```
günther<>i<>helga<>54 126 1137 110 55 54 54
g<>i<>h<>43 88 1137 66 43 43 43
i<>helga<>su<>17 1137 110 713 54 74 17
što<>sam<>rekao<>14 267 322 83 25 19 17
da<>je<>to<>12 467 1908 397 85 20 45
ivan<>i<>ana<>11 129 1137 114 16 13 12
prije<>nego<>što<>10 70 83 267 12 10 24
su<>se<>u<>10 713 992 1053 100 35 45
se<>da<>je<>10 992 467 1908 23 32 85
ivo<>andrić<>je<>10 10 11 1908 10 10 10
andrić<>je<>književnik<>10 11 1908 13 10 10 10
u<>seminaru<>je<>9 1053 52 1908 52 62 9
ana<>je<>studentica<>9 114 1908 46 34 10 17
ali<>su<>se<>9 103 713 992 15 17 100
i<>h<>su<>9 1137 66 713 43 74 9
je<>rekao<>da<>9 1908 83 467 16 41 17
su<>naši<>poznanici<>9 713 15 13 10 9 10
hrvatski<>književni<>jezik<>9 32 26 38 9 9 23
su<>günther<>i<>9 713 126 1136 9 32 55
ovo<>su<>naši<>9 72 713 15 13 9 10
bio<>kod<>kuće<>9 102 154 89 13 9 73
jeste<>li<>vidjeli<>8 44 336 32 39 8 8
bio<>u<>kinu<>8 102 1053 27 20 8 27
i<>ja<>sam<>8 1137 171 322 35 11 44
samo<>to<>je<>8 71 397 1908 8 9 67
su<>razgovarali<>o<>7 713 46 217 20 10 8
ja<>nisam<>rođak<>7 171 111 13 26 7 7
je<>u<>seminaru<>7 1908 1053 52 45 7 52
studenti<>su<>organizirali<>7 75 713 10 16 7 7
uz<>more<>su<>7 21 39 713 10 7 7
sam<>da<>je<>7 322 467 1908 14 10 85
put<>preskočio<>je<>7 43 15 1908 7 11 10
prijedlog<>nije<>prihvaćen<>7 23 274 9 7 7 8
do<>12<>sati<>7 109 10 29 7 8 7
jeste<>li<>se<>7 44 336 992 39 7 30
su<>organizirali<>štrajk<>7 713 10 22 7 7 8
se<>ne<>čudim<>7 992 420 9 26 7 8
je<>rekla<>da<>7 1908 13 467 8 41 9
vozili<>su<>se<>7 21 713 992 7 11 100
razgovarali<>su<>o<>7 46 713 217 7 9 7
kod<>kuće<>i<>6 154 89 1136 73 9 6
nitko<>nije<>odgovorio<>6 14 274 32 9 6 6
je<>li<>ovaj<>6 1908 336 70 77 16 6
rodio<>sam<>se<>6 13 322 992 6 6 34
u<>njemačkoj<>ali<>6 1053 11 103 11 9 6
kao<>što<>je<>6 116 267 1908 19 11 70
napisao<>je<>referat<>6 21 1908 19 11 6 6
gospodin<>neu<>je<>6 41 20 1908 11 13 8
studenti<>organiziraju<>štrajk<>6 75 6 22 6 6 6
otišao<>je<>kući<>6 31 1908 22 16 6 7
u<>seminaru<>koliko<>6 1053 52 120 52 15 6
njemačkoj<>ali<>su<>6 11 103 713 6 6 15
marija<>ne<>pozna<>6 69 420 9 7 6 7
su<>g<>i<>6 713 88 1136 6 32 43
i<>mi<>smo<>6 1137 141 165 19 13 20
i<>ana<>su<>6 1137 114 713 12 74 6
```

```
ana<>je<>pitala<>6 114 1908 13 34 6 9
u<>seminaru<>su<>6 1053 52 713 52 13 6
na<>prošlom<>sastanku<>6 409 6 16 6 7 6
nije<>bila<>na<>6 274 71 409 9 8 9
bili<>u<>zagrebu<>6 62 1053 35 19 6 32
ne<>bih<>to<>6 420 65 397 13 9 9
gdje<>se<>nalazi<>6 127 992 22 13 6 12
koliko<>je<>sati<>6 120 1908 29 49 7 6
je<>dva<>metra<>6 1908 59 6 6 6 6
mislim<>da<>je<>6 16 467 1908 14 6 85
je<>li<>on<>6 1908 336 140 77 18 8
hvala<>dobro<>sam<>6 23 103 322 11 7 9
i<>vi<>ste<>6 1137 89 100 15 7 22
preskočio<>je<>dva<>6 15 1908 59 10 6 6
popodne<>kod<>kuće<>6 22 154 89 6 6 73
bio<>u<>menzi<>6 102 1053 15 20 6 15
jeste<>li<>vi<>6 44 336 89 39 9 6
profesor<>i<>doktorica<>6 42 1137 16 6 7 9
su<>se<>brzo<>6 713 992 43 100 9 12
kao<>da<>je<>5 116 467 1908 19 11 85
ja<>sam<>dobro<>5 171 322 103 44 7 7
vozi<>u<>centar<>5 22 1053 14 6 5 12
je<>li<>gospodin<>5 1908 336 41 77 10 5
je<>ograničena<>u<>5 1908 10 1053 5 81 6
da<>je<>rođen<>5 467 1908 8 85 5 6
gospođica<>weber<>je<>5 17 19 1908 11 6 5
u<>kojoj<>ulici<>5 1053 28 21 15 15 5
je<>studentski<>dom<>5 1908 10 17 5 5 10
je<>pitala<>ivana<>5 1908 13 27 9 7 6
nego<>što<>smo<>5 83 267 165 24 5 10
je<>star<>profesor<>5 1908 34 42 17 9 5
su<>kod<>kuće<>5 713 154 89 5 5 73
gdje<>je<>studentski<>5 127 1908 10 59 5 5
studentu<>nitko<>nije<>5 8 14 274 5 5 9
gospodin<>braun<>je<>5 41 22 1908 15 13 5
automobil<>je<>skup<>5 38 1908 11 13 5 6
a<>nije<>se<>5 262 274 992 11 19 7
na<>predstavi<>je<>5 409 6 1908 5 31 5
još<>samo<>to<>5 124 71 397 5 5 8
je<>ostalo<>od<>5 1908 9 255 6 23 5
to<>su<>dobili<>5 397 713 20 9 5 6
brzina<>je<>ograničena<>5 15 1908 10 5 8 5
li<>stanovati<>u<>5 336 8 1053 5 17 6
```

Prilog 6: Rezultati analize 4-grama udžbeničkoga korpusa s udžbenicima KA i KF. Prikazano je prvih 100 najčešćih kolokacija nelematiziranoga teksta.

```
günther<>i<>helga<>su<>17 126 1137 110 713 55 54 17 54 74 17 54 17 17 17
ivo<>andrić<>je<>književnik<>10 10 11 1908 13 10 10 10 10 10 10 10 10 10
10
g<>i<>h<>su<>9 88 1137 66 713 43 43 9 43 74 9 43 9 9 9
ovo<>su<>naši<>poznanici<>9 72 713 15 13 13 9 9 10 9 10 9 9 9 9
su<>günther<>i<>helga<>9 713 126 1136 110 9 32 10 55 54 54 9 9 9 54
studenti<>su<>organizirali<>štrajk<>7 75 713 10 22 16 7 7 7 7 8 7 7 7 7
put<>preskočio<>je<>dva<>6 43 15 1908 59 7 11 6 10 6 6 7 6 6 6
ali<>su<>se<>brzo<>6 103 713 992 43 15 17 7 100 9 12 9 6 7 6
vozili<>su<>se<>u<>6 21 713 992 1053 7 11 6 100 35 45 7 6 6 10
preskočio<>je<>dva<>metra<>6 15 1908 59 6 10 6 6 6 6 6 6 6 6 6
ivan<>i<>ana<>su<>6 129 1137 114 713 16 13 8 12 74 6 11 7 6 6
su<>g<>i<>h<>6 713 88 1136 66 6 32 6 43 43 43 6 6 6 43
u<>njemačkoj<>ali<>su<>6 1053 11 103 713 11 9 33 6 6 15 6 6 6 6
sve<>što<>sam<>rekao<>5 103 267 322 83 7 7 6 25 19 17 5 5 5 14
uopće<>se<>ne<>čudim<>5 17 992 420 9 5 5 5 26 7 8 5 5 5 7
brzina<>je<>ograničena<>u<>5 15 1908 10 1053 5 8 6 5 81 6 5 5 6 5
još<>samo<>to<>je<>5 124 71 397 1908 5 5 8 8 9 67 5 5 5 8
studentu<>nitko<>nije<>odgovorio<>5 8 14 274 32 5 5 5 9 6 6 5 5 5 6
to<>su<>dobili<>od<>5 397 713 20 255 9 5 30 6 11 5 5 5 5 5
u<>seminaru<>koliko<>je<>5 1053 52 120 1908 52 15 100 6 6 49 6 6 13 5
da<>je<>to<>dobro<>5 467 1908 397 103 85 20 8 45 9 7 12 5 6 5
ne<>želim<>razgovarati<>o<>5 420 8 12 217 5 8 20 5 5 8 5 5 8 5
to<>je<>ostalo<>od<>5 397 1908 9 255 67 5 30 6 23 5 5 12 5 5
da<>je<>rođen<>u<>5 467 1908 8 1053 85 5 19 6 81 5 5 5 5 5
svi<>su<>uglavnom<>zadovoljni<>5 21 713 14 23 13 5 5 6 6 5 5 5 5 5
možemo<>li<>stanovati<>u<>5 17 336 8 1053 5 5 5 5 17 6 5 5 5 5
samo<>to<>je<>ostalo<>5 71 397 1908 9 8 9 5 67 5 6 8 5 5 5
o<>čemu<>se<>radi<>5 217 27 992 20 16 7 5 12 5 6 5 5 5 5
gdje<>je<>studentski<>dom<>5 127 1908 10 17 59 5 5 5 5 10 5 5 5 5
bila<>u<>njemačkoj<>ali<>5 71 1053 11 103 19 5 5 11 9 6 5 5 5 6
njemačkoj<>ali<>su<>se<>5 11 103 713 992 6 6 5 15 17 100 6 5 5 9
što<>sam<>rekao<>ne<>4 267 322 83 420 25 19 6 17 6 5 14 4 4 4
kojem<>peronu<>stoji<>vlak<>4 33 6 28 18 4 4 4 5 5 5 4 4 4 5
ana<>je<>pitala<>ivana<>4 114 1908 13 27 34 6 6 9 7 6 6 6 4 5
upoznao<>je<>nove<>prijatelje<>4 10 1908 9 6 4 4 4 5 5 4 4 4 4 4
želim<>razgovarati<>o<>vašim<>4 8 12 217 4 5 5 4 8 4 4 5 4 4 4
pogledavši<>me<>odgovorio<>je<>4 8 48 32 1908 5 4 5 4 5 11 4 4 4 4
svemu<>što<>sam<>rekao<>4 12 267 322 83 4 5 4 25 19 17 4 4 4 14
strani<>grad<>upoznao<>je<>4 13 23 10 1908 4 4 5 4 6 4 4 4 4 4
je<>ostalo<>od<>mojih<>4 1908 9 255 9 6 23 4 5 4 4 5 4 4 4
g<>i<>h<>ovaj<>4 88 1137 66 70 43 43 5 43 7 4 43 4 4 4
ne<>razmišljam<>ni<>o<>4 420 6 122 217 6 23 20 4 4 6 4 4 6 4
je<>dva<>metra<>i<>4 1908 59 6 1136 6 6 43 6 6 4 6 4 4 4
stanuju<>g<>i<>h<>4 13 88 1136 66 4 4 4 43 43 43 4 4 4 43
ići<>ću<>u<>kazalište<>4 22 98 1053 18 4 4 4 5 4 8 4 4 4 4
su<>uglavnom<>zadovoljni<>mojim<>4 713 14 23 11 6 6 4 5 4 5 5 4 4 4
cijeli<>dan<>u<>uredu<>4 30 62 1053 28 26 5 4 6 4 26 5 4 4 4
vi<>ne<>shvaćate<>naše<>4 89 420 5 8 8 5 4 5 4 4 5 4 4 4
jedan<>drugome<>jedni<>drugima<>4 96 5 12 9 5 4 4 4 4 4 4 4 4 4
sjećati<>se<>sjetiti<>se<>4 5 992 7 992 4 4 5 6 34 4 4 4 4 4
li<>g<>i<>h<>4 336 88 1136 66 4 15 5 43 43 43 4 4 4 43
to<>je<>važnije<>od<>4 397 1908 4 255 67 4 30 4 23 4 4 12 4 4
na<>kojem<>peronu<>stoji<>4 409 33 6 28 11 5 4 4 4 5 4 4 4 4
studenta<>marija<>ne<>pozna<>4 20 69 420 9 4 4 4 7 6 7 4 4 4 6
polja<>i<>šume<>su<>4 11 1137 13 713 4 4 4 4 74 7 4 4 4 4
```

ostalo<>od<>mojih<>starih<>4 9 255 9 6 5 4 4 4 4 4 4 4 4 4
ću<>u<>kazalište<>ako<>4 98 1053 18 77 5 4 4 8 7 4 4 4 4 4
s<>našom<>i<>vašom<>4 209 4 1136 5 4 15 4 4 4 4 4 4 4 4
se<>ne<>čudim<>tvojim<>4 991 420 9 10 26 7 6 8 4 4 7 4 4 4
koji<>tramvaj<>vozi<>u<>4 97 23 22 1053 4 4 5 5 5 6 4 4 4 5
su<>da<>je<>rođen<>4 713 467 1908 8 9 6 4 85 5 6 4 4 4 5
bio<>kod<>kuće<>i<>4 102 154 89 1136 13 9 7 73 9 6 9 5 4 6
ne<>čudim<>tvojim<>neobičnim<>4 420 9 10 4 8 4 4 4 4 4 4 4 4 4
to<>sam<>čuo<>od<>4 397 322 17 255 10 9 30 4 11 7 4 8 7 4
grad<>upoznao<>je<>nove<>4 23 10 1908 9 4 6 4 4 4 5 4 4 4 4
ne<>vjerujem<>ni<>u<>4 420 9 122 1053 8 23 13 4 4 9 4 4 4 4
uglavnom<>zadovoljni<>mojim<>i<>4 14 23 11 1136 5 4 5 5 5 4 4 4 4 4
i<>mi<>bismo<>se<>4 1137 141 33 992 19 4 32 4 5 14 4 5 4 4
tramvaj<>vozi<>u<>centar<>4 23 22 1053 14 5 5 4 6 5 12 5 4 4 5
zadovoljni<>mojim<>i<>tvojim<>4 23 11 1136 10 5 5 4 4 4 4 4 4 4 4
dugo<>su<>razgovarali<>o<>4 101 713 46 217 12 5 4 20 10 8 4 4 4 7
u<>strani<>grad<>upoznao<>4 1053 13 23 10 4 6 4 4 4 4 4 4 4 4
saznali<>su<>da<>je<>4 16 713 467 1908 4 4 4 9 6 85 4 4 4 4
ja<>nisam<>bio<>u<>4 171 111 102 1053 26 6 9 5 8 20 4 5 6 4
neu<>je<>star<>profesor<>4 20 1908 34 42 8 5 5 17 9 5 5 4 4 5
gospodin<>braun<>je<>student<>4 41 22 1908 94 15 13 5 5 5 19 5 5 4 4
cijeli<>dan<>kod<>kuće<>4 30 62 154 89 26 4 4 4 4 73 4 4 4 4
studentica<>nije<>bila<>na<>4 46 274 71 409 4 4 4 9 8 9 4 4 4 6
to<>ćete<>naći<>u<>4 397 20 19 1053 4 4 7 6 4 4 4 4 4 4
ima<>li<>mjesta<>u<>4 60 336 33 1053 9 4 5 5 17 8 4 4 4 4
odgovorio<>je<>ne<>pogledavši<>3 32 1908 420 8 11 5 3 8 3 4 3 3 3 3
još<>nisam<>vidio<>tako<>3 124 111 52 57 11 4 3 7 3 3 3 3 3 3
vratit<>ću<>se<>u<>3 4 98 992 1053 3 4 3 18 6 45 3 3 3 3
popodne<>kod<>kuće<>ali<>3 22 154 89 103 6 6 3 73 3 3 6 3 3 3
uz<>more<>su<>lijepe<>3 21 39 713 3 10 7 3 7 3 3 7 3 3 3
i<>petar<>su<>poznavali<>3 1137 48 713 3 12 74 3 3 3 3 3 3 3 3
odgovorio<>je<>a<>nije<>3 32 1908 262 274 11 3 3 5 10 11 3 3 3 4
ne<>želi<>razgovarati<>o<>3 420 12 12 217 3 8 20 3 3 8 3 3 8 3
l<>na<>kraju<>sloga<>3 3 409 6 3 3 3 3 5 3 3 3 3 3 3
ona<>mi<>je<>pisala<>3 112 141 1908 10 3 5 3 20 3 3 3 3 3 3
rođak<>ja<>nisam<>rođak<>3 13 171 111 13 4 4 3 26 7 7 4 3 3 7
će<>loše<>što<>god<>3 155 22 267 31 3 3 3 3 3 6 3 3 3 3
kod<>prijatelja<>su<>ostali<>3 154 22 713 5 4 3 3 3 3 3 3 3 3 3
došavši<>u<>strani<>grad<>3 3 1053 13 23 3 3 3 4 6 4 3 3 3 4
dogodilo<>uplašili<>su<>se<>3 9 3 713 992 3 4 3 3 3 100 3 3 3 3
mene<>to<>ne<>zanima<>3 32 397 420 12 4 7 3 15 5 7 3 3 3 5
studentici<>nisam<>dao<>knjigu<>3 5 111 14 64 3 3 3 3 5 4 3 3 3 3
student<>govori<>o<>štrajku<>3 94 41 217 6 3 3 3 11 5 6 3 3 3 5
sam<>bio<>kod<>kuće<>3 322 102 154 89 11 6 5 13 9 73 4 3 5 9
su<>učinili<>pobjegli<>su<>3 713 10 3 713 3 3 35 3 3 3 3 3 3 3

Prilog 7: Izvadak iz HUMOR-ove pradigme glagola *lagati*

```
0 lagati
1(1)gati(1)       gati       b 01111110 0110.... inf
1(1)žem(1)        žem        c 01111110 .110.... 1
1(1)žeš(1)        žeš        d 01111110 .110.... 2
1(1)že(1)         že         d 01111110 .110.... 3
1(1)žemo(1)       žemo       d 01111110 .110.... p1
1(1)žete(1)       žete       d 01111110 .110.... p2
1(1)žu(1)         žu         e 01111110 .110.... p3
0
1(1)gao(1)        gao        f 01111110 0110.... m
1(1)galo(1)       galo       g 01111110 0110.... n
1(1)gala(1)       gala       g 01111110 0110.... fpn
1(1)gali(1)       gali       g 01111110 0110.... pm
1(1)gale(1)       gale       g 01111110 0110.... pf
0
1(1)gah(1)        gah        h 01111110 0110.... a1
1(1)ga(1)         ga         h 01111110 0110.... a23
1(1)gasmo(1)      gasmo      h 01111110 0110.... ap1
1(1)gaste(1)      gaste      h 01111110 0110.... ap2
1(1)gaše(1)       gaše       h 01111110 0110.... ap3
0
1(1)gah(1)        gah        i 01111110 0110.... im1
1(1)gaše(1)       gaše       i 01111110 0110.... im23
1(1)gasmo(1)      gasmo      i 01111110 0110.... imp1
1(1)gaste(1)      gaste      i 01111110 0110.... imp2
1(1)gahu(1)       gahu       i 01111110 0110.... imp3
0
1(1)ži(1)         ži         k 01111110 .110.... i2
1(1)žimo(1)       žimo       l 01111110 .110.... ip1
1(1)žite(1)       žite       m 01111110 .110.... ip2
0
1(1)žući(1)       žući       n 01111110 .110.... gps
0
1(1)gavši(1)      gavši      o 01111110 0110.... gpp
0
1(1)gat(1)        gat        p 01111110 0110.... f
0
1(1)gan(1)        gan        q 01111110 0110.... gpp01
1(1)gana(1)       gana       q 01111110 0110.... gpp02
1(1)ganu(1)       ganu       q 01111110 0110.... gpp03
1(1)ganim(1)      ganim      q 01111110 0110.... gpp04
1(1)gani(1)       gani       q 01111110 0110.... gpp05
1(1)ganog(1)      ganog      q 01111110 0110.... gpp06
1(1)ganoga(1)     ganoga     q 01111110 0110.... gpp07
1(1)ganom(1)      ganom      q 01111110 0110.... gpp08
1(1)ganome(1)     ganome     q 01111110 0110.... gpp09
1(1)ganomu(1)     ganomu     q 01111110 0110.... gpp10
1(1)gane(1)       gane       q 01111110 0110.... gpp11
1(1)ganoj(1)      ganoj      q 01111110 0110.... gpp12
1(1)gano(1)       gano       q 01111110 0110.... gpp13
1(1)ganih(1)      ganih      q 01111110 0110.... gpp14
1(1)ganima(1)     ganima     q 01111110 0110.... gpp15
0
1(1)ganje(1)      ganje      r 01111110 0110.... g11;41;51
1(1)ganja(1)      ganja      r 01111110 0110.... g21;12;22;42;52
1(1)ganju(1)      ganju      r 01111110 0110.... g31;61
1(1)ganjem(1)     ganjem     r 01111110 0110.... g71
1(1)ganjima(1)    ganjima    r 01111110 0110.... g32;62;72
```